Analysis of complex socio-economical systems

PhD Thesis

Eszter Bokányi

Supervisor: Gábor Vattay, DSc
Department of Physics of Complex Systems
Faculty of Science
Eötvös Loránd University, Budapest, Hungary

Doctoral School of Physics
Head of School: Jenő Gubicza, DSc

Doctoral Program for Statistical Physics
Biological Physics and Physics of Quantum Systems
Head of Program: Jenő Kürti, DSc

Acknowledgements

First and foremost, I would like to thank my supervisor, Gábor Vattay, for his guidance and support through all of my PhD years. I am very grateful that he encouraged me to develop my skills and my own ideas, and that he trusted me with great independence. It meant a great deal to me that I could turn to him with all kinds of questions and that I could always count on his advice.

Second, I'd like to thank my collaborator Dani Kondor, whose enormous technical knowledge and kind suggestions helped immensely to start my PhD years, and with whom it is always a pleasure to discuss any new project or topic. Special thanks to Zsófi Kallus, István Gódor, Peti Kersch, and all of my former colleagues from my internship at Ericsson. Without them, I would never have been at so much ease with my own computer. I'm extremely grateful to them for passing on their enthusiasm for data, coding and new technologies. I'm also indebted to Laci Dobos, who made sure that the Twitter database survived until I finished my thesis, and who taught me how to squeeze out results in a couple of hours instead of a couple of weeks from terabytes of data.

I'm grateful for the useful comments and remarks and the help in the proofreading of various parts of the manuscript for Dani Kondor, Balázs Lengyel, and Ancsa Hannák.

I'd also like to thank my office mates Orsi Pipek, Anna Horváth and Sanyi Juhász for supporting me during the four years and throughout the writing process, and for making the time worth spending in the office.

Contents

Introduction

Most systems surrounding us in everyday life consist of the collection of numerous elements and their interactions. Tthe interactions between different parts of such systems at their smallest scale often lead to unexpected outcomes at the observable level. This kind of emergent behavior is a hallmark of *complex systems*, whose patterns may be hard to predict from the properties of their constituents. Some universal properties common to various different systems might also characterize these patterns. Understanding this universal behavior is a great theoretical challenge, while in many cases, it is also crucial for practical applications.

The tools for the modeling and investigation of complex systems were mostly developed in the field of physics. Theories and numerical simulations for self-organization and pattern formation explained crystal growth with different boundary conditions, as well as phase transitions in advanced materials. However, complex systems also include living cells, the climate of the Earth, transportation and communication systems or the economic market. This demonstrates that the examples span across multiple scientific disciplines. Therefore, the methods and concepts of complexity science can be applied to a wide variety of fields ranging from language, society, and economics to biology or machine behavior.

Previously, the lack of computing capacity and the inability to obtain and process large datasets prevented the detailed exploration and simulation of such systems. Recently, there is growing accessibility of digitalized data also in fields that are not directly related to physics. Gene sequencing and gene expression, sensor systems, internet traffic, stock market transactions all provide a large amount of information for further analysis. This enables the building and testing of new models for collective emerging phenomena.

The currently available digital data sources give an unprecedented insight into collective human behavior. Mobile phones and subscriptions are now almost ubiquitous, with a growing share of smartphones on the market. The direct usage of call data makes it possible to get a large-scale and fine-grained picture of the spatial

and temporal aspects of human activity. With the help of the internet, billions of people use various online platforms for ordering food, different items, hiring people or cars for jobs or maintaining their personal relationships through social media platforms. When people post messages, upload photos, recommend places, search for or engage with news, they leave valuable traces of information containing rich context behind. Harvesting this information is the key to building successful models of emergent patterns in human behavior, that often rely on previous concepts from physics and complexity sciences.

These digital human footprints often contain sensitive data. For example, call records or credit card transactions are able to reveal information on the whereabouts and habits of individual users. Yet, after careful anonymization, encryption and aggregation, it is possible to obtain meaningful insights into human behavior without violating these limitations.

The growing amount of high-resolution data makes it lucrative to search for patterns and make predictions without leaning on modeling. However, our understanding of socio-economical phenomena cannot come solely from data science that often lacks structural insights and context and fails to explain the detected patterns. Therefore, new approaches should develop models that bridge the microscopic data to macroscopic observables. These methods should create solid validation, calibration, and comparison of the developed models with ground-truth data.

In this thesis, I would like to contribute to the bridging of human digital data footprints with real-world outcomes through different models. In Chapter 1, I aspire to introduce the literature on how human digital footprints can be used to model various real-world outcomes, especially by using data with geographical information. The background review is then narrowed down to the introduction of the Twitter social network and the theory of urban scaling, since Twitter data and the urban scaling methodology form the background of the following chapters of the thesis.

In Chapter 2, I present the results of a study on urban scaling in historical and recent presidential elections of the United States. Here, I show that election results fit into the urban scaling framework and that a probabilistic model from economic complexity theory underlies the scaling results. In Chapter 3, I explore how the

same urban scaling phenomenon is present in the words of the language we use on social media, namely the Twitter online social network, and how qualitative linguistic laws such as Zipf's law and the Heaps law hold in these online posts.

Chapter 4 analyzes the correlation of employment and unemployment rates of geographical areas in the United States with Twitter daily activity profiles. In this chapter, I also develop an algebraic approach for treating geographical areas by assuming that the observed human activity timelines consist of the timelines of differently behaving groups of people. Chapter 5 investigates how spatiotemporal patterns in social media word use predict mobile-phone based land use clustering in different cities. Finally, I am going to summarize my findings in Chapter 6.

1

Background

1 Predicting real-world outcomes

The fields of *computational social science* has emerged recently partly due to the shift in social sciences towards more data-intensive research and a growing computing capacity for data storage and simulations [10]. The introductory article that attempted to summarize existing and foster future research under the hood of this field came out in 2009 [11]. Some large-scale social science measurements have been in use decades before the appearance of this paper. But the sheer amount of new data sources and the possibilities that come with them encouraged the development and testing of new techniques and their limits [12].

In the following, I attempt to give an overview of the different possibilities and limitations of the field of computational social science, with a special focus on connecting digital geographical footprints to real-world data or outcomes, and with a predominance of literature on the Twitter social media platform.

Using data from online sources to predict real-world outcomes is an alluring topic, that also has remarkable economic potential. Most companies aspire for gathering information on user preferences concerning a product or a service, or predicting sales before releasing an item. Therefore, the paper of Asur and Huberman [13] earned considerable notice. The authors predicted future movie box-office revenues

better than market-based predictors. They have successfully built the prediction model using messages of an online social networking site, Twitter. By incorporating sentiments extracted from the short messages of Twitter into their model, they could further improve on the performance of their predictions. The paper claimed that "[their] method can be extended to a large panoply of topics, ranging from the future rating of products to agenda setting and election outcomes. At a deeper level, this work shows how social media expresses a collective wisdom which, when properly tapped, can yield an extremely powerful and accurate indicator of future outcomes." Though this claim might seem too dashing at first sight, with different tools and careful approach, it is possible to harvest the rich source of information coming from similar data sources.

Search engines are one of the first tools users consult when they are confronted with a new question. What people search for often reveals geographical or temporal trends, for example when they gather information on political candidates or a beginning illness. One of the first articles that assessed the modeling of the geographical aspects of search engine queries was the article by Backstrom et al. [14]. They determined the rough location of search engine queries by estimating the location of the IP-address of the originating query and then fitted a model that predicted an approximate center and influence distance for a certain query. By comparing the results of the study to ground-truth data, the results suggest that online content is very much embedded in real-world geographic context. This means that online content may contain information that is strongly related to socio-economical phenomena having spatial aspects. Another paper of the same group [15] proposed a method for inferring the location of Facebook users from the location of their friends by using a similar probabilistic approach. This leads to a location precision comparable to that of IP-based methods.

A famous use case of using search engine queries for real-world predictions is the Google Flu Trends tool for influenza epidemic forecasting [16]. At first, the method has gained much publicity, since the authors claimed that the tool is able to follow current influenza epidemic trends from symptom search volumes on Google with only a one day lag. In contrast, the official organization that collects data reported by healthcare professionals, Center for Disease Control (CDC), has much slower reaction times. Such methods could be very useful for reducing public health

costs because early forecasts could mean that effective preventive measures such as vaccination can be taken at an early stage of the epidemic. In 2013, though, Google Flu Trends again made the headlines, for predicting more than double influenza-related doctor visits than CDC. The article [17] stressed that the algorithm performed poorly because of overfitting a few numbers of parameters from a large dataset. By using two- or three weeks of historical CDC data, predictions are actually better than that of Google Flu Trends. Moreover, the search term suggestion feature introduced into Google search in 2011 modifies the query term distribution because of the reinforcing effects of the recommendations, therefore, it influences the results of the predictions. The moral of the case of Google Flu Trends is that it is very important to reflect on the generalizability and the context of the problem we are currently trying to solve using non-traditional data sources.

Mining public health indicators from social media remains a promising field despite the shortcomings of the Google Flu Trends predictions. It takes time for the official reports to be released. Meanwhile, self-reported influenza-like illnesses or hay-fever symptoms from online data sources can mark the beginning of the disease or allergy season almost instantaneously. Using Twitter messages containing GPS information, the so-called *geolocated tweets*, it is even possible to pinpoint the location of illnesses or allergens, that is very useful for creating timely and local interventions [18]. Several studies explored the possibilities of automatically constructing health-related topics from Twitter and various other online sources to complement traditional reports [19–15].

Because different psychological traits are present in the language with which users express themselves [25–19], another direction in connecting public health to social media research is that of identifying the potential risk, course and onset of mental illnesses [29, 21]. The online language reflects not just on the psychological state of individuals, but also that of a community. Thus, the rate of coronary heart disease failures correlates with the linguistic traits of social media posts representing negative mental states [31]. Not just the diseases, but their perception can also be tracked using social media. For example, discussions on vaccines [32], diabetes expertise, news, and management [33] or Ebola misinformation and rumors [34] can be assessed by working with appropriately filtered Twitter data.

Another huge problem is the prevalence of obesity in developed countries, that leads to several different civilization diseases in the population. Therefore, it could be helpful to assess information on dietary habits of users and peer effects in food consumption using Twitter messages at a large scale [35, 27]. Obesity rates in the United States [37, 29] have been shown to be correlated with Twitter-derived characteristics such as happiness, food, and physical activity indicators at the zip code level.

Using geolocated social media posts to locate pollen seasons reflects an engineering attitude towards the data. In this engineering framework, humans are regarded as a distributed sensor system, and posts containing information relevant to the selected topic are considered as noisy signals. These noisy signals are then mined for meaningful information and are considered as an "alert" if the meaningful part surpasses a certain threshold. A very early article of Eagle and Pentland emphasizes the role of mobile phones as new, wearable sensors [39]. All kinds of alerts can later be built on this sensor network [40] after processing the sensor information. For example, in an emergency case, users messaging about the urgent situation may reach locals faster than official emergency centers. Japanese authors report a Twitter-based earthquake-alert system that is able to reach users potentially at risk at a higher speed than official alert systems [41]. The data footprints of moving disasters and their consequences such as tornadoes are also traceable through geolocated posts, indicating the merge of the real and the cyberspace [42–35].

Geolocated social media data also means that the physical space is enriched by the content from the online space through the different pictures, posts, and metadata that are generated on the different platforms. This phenomenon can be thought of as some kind of an augmented reality, in which user-generated content layers, though invisible to the eye, add a considerable amount of information on top of traditional maps [45]. However, this content emphasizes only selected elements of the social experience, for example, the online representation of languages can be distorted compared to the number of speakers. Moreover, recommendation algorithms presenting results for queries can cause a further imbalance by reinforcing the visibility of already highly popular viewpoints, languages, and places.

Therefore, the careful evaluation of the social composition of the creators of these contents is necessary.

This is why many studies deal with predicting the demographical traits of social media users including Twitter users. Users' self-reported first and last names often serve as a predictor of race and gender. Because certain names have a limited popularity time span, they can also serve as a predictor of age [46–40]. Geographical location of user homes, which can be determined for frequent posters, can be connected to census data, thus inferring demographic attributes for social media poster populations [50]. Follower lists and content can also be indicative of user demographics, according to an article [51], in which the authors predict demographical attributes using the census information on visitors of different websites combined with the follower information. Age might also be encoded in linguistic style, the proportion of certain word categories and the content sharing habits of users. Based on these traits, a machine classifier predicts the age of users similarly or better, but certainly faster than humans [52]. Machine learning algorithms are able to classify users not just into demographic, but also into political categories [53, 45].

If not inferring attributes for individual users, correlating or predicting socio-economic variables of geographical areas from geolocated data footprints is also a widely researched topic. Social network structure, activity, and language are all influenced by the ethnicity, age, gender, social status or lifestyle of users. Thus, they allow for various methods to search for connections or models between the online content and the demographic or economic variables. The diversity of social networks is also strongly related to the economic output of geographical areas [55]. Distinct behavioral patterns characterize rural and urban dwellers in their mobile communication patterns, and signs of adaptation are also traceable as people move to more urbanized spaces [56]. Local social network structure can be linked to corruption [57], and user geolocation allows the studying of long-distance traveling and migration patterns on a previously unprecedented scale [58, 50]. Main language use patterns uncovered by unsupervised learning methods are also strongly correlated to county-level census variables in the United States [60]. Not just English, but also French linguistic features show a strong dependence on socio-economical

factor [61]. Social media posts also reveal immigrant communities and their level of integration or segregation within cities [62, 54].

2 The Twitter social media platform

Twitter is an online social networking site that has been launched in 2009. With this service, users can post short messages, so-called *tweets* to other users that subscribe to these messages. The service advertises itself as a so-called microblogging platform, where the term microblogging refers to a limit on the length of tweets. While traditional blog posts are typically refreshed on a daily scale, Twitter encourages users to broadcast their short status updates, the so-called *tweets* more often. The limit on the character number of a tweet used to be 140, but it has now been extended to 280 characters. The relative simplicity of the concept of the service attracted a great many users, with the number of monthly active tweeters estimated to be above 300 million in each quarter since 2015 according to the company's own report [64].

Messages can not only contain plain text, they might contain a variety of hypertext elements. They often include images and web links, many of the latter being abbreviated by various services due to the 140/280 character constraint. Users can mention another user in their tweets by using the `@` sign and the *screen name* or username of the mentioned user in the text. They can also mark tweets as related to certain topics by highlighting certain words with a `#` sign, that is used to create so-called *hashtags*. Each tweet has a timestamp that bears the exact UTC millisecond time of the posting. Conversations are created by replying to a tweet: the original tweets are displayed in the same thread with the replies on the webpage or in the mobile application.

The user that subscribes to another user's stream of tweets – or the so-called *feed* – is called a *follower*. Followership is an asymmetrical relationship since interest in others' content is not implicitly mutual. A typical example of such an asymmetrical relationship is the Twitter account of a news outlet, that is followed by much more people that it follows. Therefore, simple follower relationships create a directed network, in which nodes are the Twitter users, and only mutual followerships form

Figure 1.1. A sample tweet from the official account of Twitter. At the top, there is a small profile picture of the user who posted, the screen name **Twitter** in bold, and the username `Twitter` beginning with . There is a button called "Follow" in the top right corner. The timestamp of the tweet is indicated below the main text. At the bottom, several counters show the different interactions with this tweet. These are the number of retweets, the number of likes and the number of replies in the conversation.

reciprocated edges in both directions. These bidirectional edges are then supposed to mark a stronger relationship between users, as such, they are often called *friendships*. Many applications only consider these mutual connections for constructing a social network from Twitter data, but usage might depend on the context.

The feature that Twitter is really famous for is *retweeting*, which is when users choose to broadcast another user's message to their own followers extended by their own comment. This enables the almost instantaneous amplification of news or messages that somehow grab the attention of many users. Therefore, the site is widely used by celebrities and politicians, since millions of followers and the followers of these followers etc. are easily reached within seconds with the right tweeting technique. This kind of instantaneousness means that news and opinions travel on Twitter much faster than they used to spread via traditional mass media such as newspapers or television broadcasting. Therefore, retweeting transformed the production and consumption of news in a profound way [65]. Also, the article [66] suggests that Twitter behaves more like a new media in many aspects rather than a social network. From studying the follower network, the authors

suggest that the network characteristics deviate significantly from that of other human social networks, since the degree distribution does not follow a power-law, and the network has a short effective diameter and low reciprocity. The power-law assumption breaks down because there are many people that gather more followers than a power-law tail would predict. These are celebrities or politicians who actively engage with their followers on the platform. However, other studies confirm some similarities between the network srtuctures derived from Twitter data and other human social networks. For example, mention networks have a small-world property [67], follower networks are highly assortative [68], and scale-freeness and small-worldness is present in the retweet network [69].

Another strong side of the Twitter data is its richness in geographic information apart from the messages and the social network data. Before April 2015, the smartphone application tagged messages with the exact GPS coordinates of the users by default, these are the so-called *geolocated* tweets. Therefore, approximately 1% of all tweets sent had GPS coordinates attached to them. After April 2015, users had to opt in to share their exact coordinates, and place-tagging by using real place names was preferred by the platform. Since tweets about the Eiffel Tower sent from New York might also be tagged as Paris, this tagging procedure is not a reliable source for the origin of the tweet. Therefore, the change in the geolocation sharing policies meant a significant decline in the number of geolocated tweets. Despite the shrink in the geolocated data volume, Twitter remains a valuable source for precisely geolocated content, but the quality and quantity of geolocated data depends on the collection period.

The availability of exact geographical location combined with the network structures made it possible to test former social science theories with a big data approach. For example, a well-known result is that the capacity of the human brain limits the number of close social relationships one can engage in. This capacity is marked by the Dunbar number, that means that an average human is closely connected to only 150 other humans. A 2011 study of Goncalves and colleagues find that this limit is still there in the age of online social networks, where in theory, it became much easier to maintain relationships regardless of physical distance and time constraints. That geographical distance matters in the formation of mutual

ties, and that the digital age did not bring the "death of distance" forward is further confirmed by another article [70].

Geographical embeddedness makes it possible to repeat the famous Milgram experiment, that claimed that there is on average six degrees of separation between any two individuals, and that messages are able to find their way to the recipient along social network hops. Researchers could test this theory by navigating through the large-scale geographical social network of individuals on Twitter. It turned out that the Twitter follower network is navigable between cities as messages find their way in the follower network quickly through long distances. But this navigability breaks down inside cities, where the dependence of the existence of social ties on distance behaves differently [71]. Weighted ties between geographical regions can be used to investigate regional clusters at the country scale [72] or at a lower pixel resolution [73]. The communities found with different clustering techniques on these weighted networks uncover meaningful economic, cultural and administrative relationships. The same weighted regional network can be used to model the geographical diffusion of new concepts or ideas, such as the viral video Gangnam style [74]. The viral spreading process on the regional social network exhibits the same characteristics as usual epidemic models using the effective distance concept of [75]. References [71–65] are based on the same dataset as Chapters 3 and 4 of the present thesis.

Twitter has been made truly popular amongst researchers by making its data freely accessible through the Application Program Interface (API). Free access is granted to at most 1% of all sent messages at a given time, and continuous filtered streaming of this 1% restricted to a topic or a geographical area is also available. If the demanded data volume surpasses this 1%, Twitter starts to sample the results. For Chapters 3, 4 and 5 of this thesis, I use the collected results of the Twitter Streaming API filtered for geolocation information. Chapters 3 and 4 rely on a large relational database containing the geolocated world stream from 2013 to 2015 [76], and Chapter 5 is based on several smaller streams confined to the geographical area of three cities.

2.1 Potential limitations and biases

While Twitter offers an unprecedented amount of freely available social media data, the dataset also has its limitations and possible biases. First of all, the algorithm with which Twitter provides the 1% sample through its Streaming API is not publicly available. Therefore, the freely available tweets might not be uniformly sampled from the Firehose stream that contains the full amount of tweets. Unfortunately, the Firehose stream is expensive to obtain and the amount of downloaded data is costly to assist with infrastructure.

To address this problem, Morstatter et al. designed comparisons between the free sampling API and the full Firehose stream in two studies. In their earlier article [77], the authors showed that the coverage of the Streaming API depends on the overall volume of the tweets matching the filter conditions. It means that filtering for certain keywords might cause a decrease in the coverage if the keywords get more global attention, or if the total volume of tweets decreases causing the lowering of the 1% threshold. However, if queries are specific enough, it is possible to obtain sufficiently large coverages. For example, since geolocated tweets amount to only a small fraction of the total tweet volume, filtering for them can yield coverages up to 90%. Also, datasets obtained with the Streaming API are consistent in the sense that the same queries give the same results if their scopes overlap in time, even when submitted from different geographical locations, which is also confirmed in [78]. The second article [79] addresses the difference in the sampled and the full stream in top hashtags, topic distributions, network metrics, and geographical distributions. They come to the conclusion that a random sample of the same size as the sample from the Streaming API represents the full stream better in the topic distribution, the top-n hashtags when n is small, and the network metrics, but the results can be enhanced by aggregating several days' worth of data.

An alarming issue is that the sampling algorithm of Twitter has not been released publicly. Whenever the volume of the data stream surpasses the 1% threshold, the API behaves like an unavoidable "black box", of which researchers could not know what biases it could introduce. A recent paper from Pfeffer et al. successfully reverse engineered the sampling algorithm [80] and showed that the sampling is based on a certain millisecond-level time-window. Tweets falling into this time

window are going to be included in the sample stream. The decision is based on the timestamp a tweet receives when arriving at Twitter's servers. Thus, if the time delay between the posting computer and Twitter's servers can be estimated reliably, it is possible to artificially enrich the supposedly random stream coming from the API. The authors in [80] demonstrated their ability to tamper with the sample, and proposed methods for identifying possible accounts that deliberately distort the sample. The found accounts are mostly automated users, so-called bots. The messages of these automated accounts are either filtered in our work, or we use data from geographical regions where tweeting activity is high, and therefore, the fraction of bot messages compared to the whole sample is negligible. Moreover, less bots post messages with coordinates than without.

Aggregation and using only geotagged tweets ensures that the data used in the present thesis contains most messages from the full stream. However, the fact remains that Twitter users are not representative of the whole population. In the United States, urban, young (aged 18 to 29) and African American users are overrepresented on the platform according to a Pew Research Survey [81]. This is in line with another piece of research showing that volunteered geographic information such as Twitter, Flickr or Foursquare (the latter two being other location-based social networking sites) is biased towards urban perspectives [82] in the US and that Twitter is more likely to be adopted by African Americans among young users [83]. Automated methods also confirm some of these biases [46], namely that Twitter users significantly overrepresent the densely populated regions of the US and are predominantly male. However, this particular study argues that geographically, Hispanic and African American people are underrepresented in counties that traditionally have higher percentages of these ethnicities (e.g. Southwest of the US for Hispanic, and Midwest and South for African American users). Highly disproportionate spatial distribution of geotagged messages in London can be attributed to the abundance of tourists tweeting from the city center, which can be another source of error in the sample [84]. In the UK, Twitter users also tend to be much younger than the population in addition to certain occupations being overrepresented on the platform [85].

The ability to broadcast information instantaneously to a large number of people in a simple yet effective form contributes to a large number of automated ser-

vices prevailing on Twitter. Robot tweeters announce weather broadcasts, post job hiring messages or advertisements at regular intervals with a very high frequency compared to that of an average user. These messages can distort samples where researchers' main aim would be to detect patterns of human interactions and conversations in Twitter messages. Therefore, automated bot detection methods are being developed to sufficiently filter bot content [86, 78]. Another topic that has attracted much attention lately is that of fake news. Fake news can also be disseminated automatically, that might influence the opinion formation processes taking place on online social platforms. Combatting disinformation through the understanding and effective filtering of bot content is crucial to maintaining the fairness of media platforms during elections or referendums [88, 80].

Therefore, when designing studies and interpreting results, we have to be aware of the sampling and demographic biases, as well as possible automated sources distorting the sample. Geotagged tweets are recovered almost fully using only the Streaming API, therefore, sampling biases do not play a large role in the datasets of the thesis. Demographic biases may, on the other hand, affect the outcomes of the models.

3 The theory of urban scaling

More than half of the world's population now lives in urban areas. The fraction and the total number of the urban population are both going to increase in the next decades [90]. Therefore, an important aspect of the utilization of data with geographical information is to understand how to create sustainable cities. This includes the quantitative analysis and modeling of the population growth, resource use, infrastructure volume, but also the economic outputs or other metrics that are needed or generated by cities.

Cities are sometimes compared to biological organisms, where the transport and infrastructure play the role of the circulatory or metabolic systems [91]. Scaling in biological systems is a long-studied subject. For example, we know that the metabolic rate of an animal depends only on the body mass with a power-law relationship that holds over several orders of magnitude [92]. This relationship

is called Kleiber's law, and it states that if M is the body mass of an animal, then the metabolic rate is proportional to $M^{3/4}$. The relationship is remarkable in its magnitude range of validity and its simplicity, as well as in the specific 3/4 exponent of the power-law. It can be shown that the exponent can be gained by supposing principles of optimal energy use and optimal material transport in all kinds of species. In this sense, the law is very general, since it assumes that a cat is just the scaled-up version of a mouse, and an elephant is just a scaled-up version of a cat in terms of energy consumption and general functioning. Similar scaling laws can be established for the length of life or for the hearbeat rates of different animals.

This idea of a generalized animal whose basic measurable quantities can be captured through universal power-laws led to the idea of cities from the same city system or country being a "scaled-up" version of each other [93, 85]. That is, between certain total input or output quantity Y of a city, and its population size N, there is a power-law relationship

$$Y(N) = Y_0 \cdot N^{\beta}, \tag{1.1}$$

where Y_0 is a constant, and the exponent β characterizes the growth of the quantity Y with population size [93]. This equation is called an *urban scaling law*, and if $\beta = 1$, it simplifies to a simple linear relationship. It turns out from measurements, that for infrastructural measures β is smaller than 1, which is called a sublinear scaling. It means that agglomerating populations into bigger and bigger cities creates economies of scale, where agglomerating people into bigger settlements pays off in having to build fewer roads, cables or gas stations per capita. This behavior is similar to that of biological organisms, despite the exact value of the exponent β being not 3/4 in the case of cities. For other quantities, though, such as the total GDP of a city, the number of patents or crime, there is a $\beta > 1$ superlinear increase with city size in the scaling law. The $\beta > 1$ superlinear behavior is remarkable since it means an increase in per capita productivity and creativity, that has no corresponding counterparts in biological systems. Therefore, it is important to understand and model the phenomena that lead to this superlinear behavior.

One of the first modeling approaches was that of Arbesman, Kleinberg and Strogatz [95]. The authors assume that output production stems from the interconnected

inhabitants of a city with outwards connections giving a negligible contribution and that the social network is organized as a random network upon a hierarchical structure. The hierarchical structure is represented as a tree with at most b branches at each level, that might – from the bottom to the top – correspond to people forming households, households forming blocks etc. They define the distance d between two people as the distance to their first common ancestor in the hierarchy. Then they suppose that the probability that two people know each other decays exponentially with d and that as d increases, the productivity being in interpersonal links varies exponentially. That is, in most cases increases, but a decrease is also imaginable. Also, they assume that the maximum number of people possibly reachable at distance d goes with b^d. By summing up these assumptions for all levels of the hierarchy for all people, productivity becomes a superlinear function, which can be modeled for example by a power-law. The superlinearity of the productivity also holds in case of looser assumptions for the decay of the friendship probability or the increase in productivity with the distance.

A similar hierarchical infrastructural approach can be found in the work of Bettencourt [96]. Here, a set of different interaction types or social networks contribute to a certain output measure Y in the city. At the same time, resources are consumed through hierarchical infrastructural networks. By assuming that infrastructure networks fill the whole area of the city, that they cover the minimum amounts consumed, and that matching boundary conditions of the hierarchy (such as each person at the lowest level needs a constant amount of energy, water etc.), it is possible to maximize the output via optimal social connections while taking energy dissipation in the network hierarchy into account. The model predicts an exponent of $\beta = 7/6$ for socio-economic outputs, and $\beta = 5/6$ for infrastructural measures. This is in line with the observed superlinear and sublinear behavior of such metrics.

A quite different viewpoint is explained in [97]. Here, the superlinearity in the exponents for socio-economic metrics is explained through the increasing economic complexity [98] of city environments with size. This model not only accounts for the superlinear scaling exponent, but gives a relationship between the constant Y_0 and β, and accounts for the behavior of the deviations around the scaling laws. Because this model underlies the model for the scaling of metropolitan area election results, I explain it in detail in Chapter 2.

An important question in this topic is the exact definition of cities for the measurements. Both the delineation and the level of aggregation can affect the measured β values. To comply with the theoretical assumptions of homogeneous population mixing and accessibility of the models, cities are defined as functional units, where strongly interacting, co-located social networks exist [99]. It can be shown that disaggregating a single unit into two smaller ones decreases the β exponents in case of superlinear scaling, if the smaller units are very heterogeneous. For example, if they represent a business and a residential part of a city, the inhomogeneity decreases agglomeration effects. This is why it is important to include all parts of a city into the delineation that play a role in either the social network or economic production. If two disparate units are counted as one, it also decreases the exponent towards linearity. In this case, the expected payoff from the agglomeration does not take place, and as a consequence, the unit is going to fall "below" the scaling law, and the slope of the fit is going to decrease [99]. In the case of the United States, both the literature and this thesis use the Metropolitan and Micropolitan Statistical Areas given by the US Census Bureau [100]. These areas are functional units created considering commuter flows and economic dependence as well as population densities, where metropolitan areas may comprise of several cities.

In a broader perspective, urban scaling laws are just the expected values of Y given the conditional probability distribution $P(Y|N)$. Estimating $P(Y|N)$ is a difficult task, though, because of the often granular nature of the data for small Y and N values. For example, the yearly number of homicides might be equal to 1, 2 or 3 for a small settlement. Therefore, by using Bayes theorem

$$P(Y|N) = \frac{P(N|Y)P(Y)}{P(N)}, \tag{1.2}$$

and the theorem of total probabilities,

$$P(N) = \sum_{Y} P(N|Y), \tag{1.3}$$

it is possible to circumvent this limitation. Thus, there will be enough number of cities for small granular values of Y.

By fitting lognormal distributions to $P(N|Y)$ for the number of homicides in three Latin-American countries, and by making a power-law assumption based on the data for $P(Y)$, the authors in [101] were able to establish a lognormal distribution for $P(Y|N)$, that is further validated in [102]. The mean value of this lognormal distribution gives the power-law of the urban scaling. Moreover, by summing up the terms in (1.3), it is possible to deduce that urban scaling leads to a Zipfian distribution $P(N)$ in the city sizes. The scaling exponents and Zipf exponents are connected, and not just the often cited $\alpha = 2$ is possible as the Zipf's law for cities (see the review of Gabaix [103]).

The exponents β and the constant Y_0 may change slowly over time, though. Therefore, scaling exponents can also be interpreted as marking different stages in the lifecycle of technologies. According to [104], new, resource-intensive, innovative technologies are the ones that are still at the first stage of the innovation lifecycle, and that are most likely to appear in bigger cities. Therefore, these technologies are characterized by a superlinear $\beta > 1$ exponent. As technologies mature and become more widespread, the exponent gradually decreases through a linear stage, where $\beta = 1$, to the sublinear $\beta < 1$ regime. This evolution is observable in the time series of the exponent of the textile industry or rubber manufacturing, that already had time to reach the sublinear state, despite having initially been superlinear processes.

Urban scaling has another effect on the mechanism of innovation cycles [93]. If we assume that urban growth is scaling-driven, that is, if resources Y are used for maintaining (with a rate R) and adding new individuals (with a rate E), than the we get that

$$Y = RN + E\frac{\mathrm{d}N}{\mathrm{d}t}, \tag{1.4}$$

and the growth equation combined with the scaling law for Y becomes

$$\frac{\mathrm{d}N}{\mathrm{d}t} = \frac{Y_0}{E}N(t)^{\beta} - \frac{R}{E}N(t). \tag{1.5}$$

The solution for the latter equation has different behavior for the superlinear, linear and sublinear cases. While the sublinear solution leads to saturation, and the linear to exponential growth, superlinear $\beta > 1$ means that cities grow faster

than exponential. This would lead to an infinite population in a finite amount of time. Thus, the creation of innovation would anticipate the collapse of cities, had it not been for resetting the initial parameters of the equation. The solution is provided by successive cycles of innovation that reset this singularity, and enable cycles of population growth spurts, which is also observable in empirical datasets.

Because the urban scaling relationships are only the expected values of the $P(N|Y)$ statistics, it is necessary to treat the deviations of the data from the scaling curves consistently. According to Alves et al. [105], the logarithmic residuals ξ have a normal distribution ($\mu = 0,\ \sigma = 1$) for all of the investigated urban indicators, where the residual ξ_i for one city having a population N_i with an urban indicator value Y_i is

$$\xi_i = \frac{\log Y_i - < \log Y_i >_w}{\sigma_w}. \tag{1.6}$$

Here, $<>_w$ denotes a window-wise average for a certain small range of N, and σ_w is the standard deviation of the Y_i-values within the same N-window. The σ_w values are independent of the population size N since they are almost constant over the whole population range for numerous different urban indicators. The deviations ξ_i from the power-law can form the basis of more reliable indicators than traditional per-capita ones. Per capita indicators introduce a bias towards smaller or bigger cities depending on whether there are sublinear or superlinear scaling laws [106]. This is why the authors of [106] call this measure so-called Scale-Adjusted Metropolitan Indicator (SAMI). The distribution of SAMI values can also be indicative of the validity of scaling laws. For example, metrics that have a skewed lognormal distribution are most likely the result of the addition of two or more power-laws [99].

These deviations ξ_i from the scaling laws tend to be persistent over time [107]. Because of the relatively long characteristic timescales, policies focusing on simple population growth might not stop undesirable persistent deviation patterns as long as there is no fundamental change in local urban dynamics. Though they would place the city into a different position along the scaling law, the under- or over-achievements would still resemble the earlier ones. On the positive side, investing in fundamental changes seems to have a long-term effect. Therefore, if successful, they could be very desirable by policy-makers. As opposed to temporal correla-

tions, spatial ones in ξ_i values last no longer than approximately 200 km. Instead of spatial proximity, cities can find their "kindred cities" by creating hierarchical clusters based on the SAMI values ξ_i. These cities may be far from each other in space, but they possess similar scale-independent characteristics. The Metropolitan and Micropolitan areas of the United States, for example, fall into a small number of categories according to this clustering as in Figure 4A-B of [107]. The concept of kindred cities also gives useful insight into the emerging urban patterns in India [108].

The theory of urban scaling has received criticism for arbitrarily choosing the power-law functional form and for the lacking statistical evidence of the lognormal residual distributions [109]. However, the proposed logarithmic relationship in [109] approaches singularity as city size decreases, and as such, it cannot account for the smaller cities in the data in many cases [99]. Also, the theoretical models proposed in [95, 97] and [96] account for the power-law scaling form. The potential bias in the OLS fits and the statistical testing of the validity of the exponents is assessed in [110] by using MLE estimators that comply with the expectations of [111–104] for power-law and power-law distribution fits. The article [102] addresses the issue of the residual distributions and the power-law distribution of $P(Y)$ used in [101] for Brazilian datasets, and finds that for most metrics, neither log-normality in the residuals, nor power-laws for the metrics distribution can be rejected, with the only exception being that of the number of homicides. Log-normality testing for the distribution of the residuals can, therefore, be used for testing whether a certain measure exhibits urban scaling [99].

2

Universal Scaling Laws in Metro Area Election Results

In this chapter I explain the anomaly of election results between large cities and rural areas in terms of urban scaling in the 1948-2016 US elections and in the 2016 EU referendum of the UK. The urban scaling curves are universal and depend on one single parameter only, which is the scaling exponent of the party that shows superlinear scaling and drives the process. The sublinear exponent of the other party is merely the consequence of probability conservation. Based on a recently developed model of urban scaling, I give a microscopic model of voter behavior. In this model, diversity characterizing humans in creative aspects is replaced by social diversity and tolerance. The model can also predict new political developments such as the fragmentation of the left and "the immigration paradox".

The material presented in this chapter appeared in [1].

1 Introduction

Based on notions from economic complexity and cultural evolution, Gomez-Lievano, Patterson-Lomba, and Hausmann recently proposed a new model (GLPLH model) of superlinear urban scaling [97]. They demonstrated the validity of the model on 43 urban phenomena related to employment, innovation, crime, education, and diseases. The model accounts for the difference in scaling exponents and average prevalence across phenomena as well as for the difference in the variance within phenomena across cities of similar size.

The central idea is that an urban phenomenon needs M number of complementary factors that must simultaneously be available for an individual to participate in an urban phenomenon. These factors might be provided by the city itself, or the individuals themselves might possess them. For example, to get a patent, at least the following six factors should be present from either of the two sources: have a technological problem, have a solution, present the idea clearly, apply for a patent, include subsequent corrections from examiners, and satisfy all the legal requirements.

There are a certain number of factors that an individual possesses by default. The probability that an individual lacks some factors that the city has to provide is chosen from the binomial distribution with probability $q \in (0, 1)$ and the number of factors M. This probability q quantifies the *complexity* of the phenomenon, because the greater q is, the more factors an individual requires from the city. The fraction of factors that a city provides for an individual is $r \in (0, 1)$. It represents a measure of urban *diversity* and tends to accumulate logarithmically $r = a + b \cdot \log N$ with the population size. Here, a and b have been found to be constant across a wide range of urban phenomena. The logarithmic accumulation of diversity is a strong assumption, but multiple anthropological studies showed that the accumulation of the diversity of skills, behaviors, beliefs or even vocabulary follows this relationship, see the references of [97]. Alternatively, the fraction of factors *not* present in a city is $1 - r = b \cdot \log N_0/N$, where $\log N_0 = (1 - a)/b$, and $N_0 \approx 1.8 \cdot 10^{14}$ is a hypothetical maximal diversity.

Given a city with m factors present, the probability that an individual requires any number of the m factors that the city has, but none of the $M - m$ factors that the city does not have is $P = (1 - q)^{M-m} \approx e^{q(M-m)}$ for $q \ll 1$. This is exactly the probability that all factors are there for a phenomenon and the individual participates in the urban process. The average number of occurrence of the phenomena is then $Y = N\langle P\rangle_N$, yielding $Y \approx Ne^{qM(1-r(N))} = Ne^{qMb\log N_0/N}$ where $\langle e^{-qm}\rangle_N \approx e^{-Mr(N)}$ and averaging goes for cities of population N. Introducing the the scaling exponent $\beta = 1 + Mbq$, this scaling curve then takes the universal form

$$Y = N_0 \left(\frac{N}{N_0}\right)^{\beta}, \tag{2.1}$$

where N is now the part of population conceivably susceptible to the given urban phenomena. The equation (2.1) now has the usual form of urban scaling, using N_0 transformed with the power $1 - \beta$ instead of Y_0 as the constant multiplier.

Scaling laws and universality have been observed in various aspects of the political process and elections [114–111]. They can be even used to detect election anomalies [121]. In the last decade, complex systems-based approaches using social contagion theory have been developed [122–119] for understanding scaling in election data. In this chapter, I concentrate on the phenomenological aspects of the observed scaling only and don't study the detailed mechanism behind the social contagion process.

In the 2016 presidential elections of the United States, it has been noticed that votes for Democrats were disproportionally high in large cities [129]. In the United Kingdom, major cities followed a similar pattern, when they mostly voted to remain in the European Union. This phenomenon can be understood in the context of social contagion, where larger cities are shown to facilitate opinion spreading due to network effect [128]. However, the exact statistical properties of the dependence of the votes on city size can be better explained through scaling laws. Here I show that election data in the US and the referendum votes in the UK show strong evidence of urban scaling. Moreover, in the US, the scaling curves follow a hidden rule, a single parameter family of scaling curves, for both parties and for all the elections in the investigated period of almost 70 years. Using the concept that tolerance and diversity are strongly coupled in cities [130], I develop a microscopic model of voter behavior which produces the macroscopic level urban scaling, explains the observed

single parameter scaling, and describes the distribution of deviations from the macroscopic curve. The new model can even explain unexpected voter behaviors like "the immigration paradox" in Britain. The paradox was that communities that had the fewest recent immigrants from the EU were the most likely in wanting to leave the EU [131].

2 Materials and methods

2.1 Data sources

First, I analyze data for the votes cast for the two main political parties in urban areas in all post-World War II US presidential elections [132] and in the UK EU referendum [133]. I downloaded county-level historical US presidential election datasets from [132], and then I calculated the total number of votes for the Democratic and Republican Party and the turnouts for all Metropolitan and Micropolitan Statistical Areas [134] by matching MSA's to the county level data [135].

As for the UK, I downloaded electorate-level number of votes for Remain and Leave from the EU referendum result dataset [133]. I filtered the UK electorates based on whether they have a city in their core [136], because the resolution of the data available about the referendum was not enough to consider using cities as units.

2.2 Data fit

For each year y in the US dataset, and also for the EU referendum data, I assume that the expected value of the number of voters for a party or an opinion (D, Democrat or R, Republican etc.) scales with the size of a city in the following way:

$$Y^{(y)}(N) = Y^{(y)}_{0,D/R} \cdot N^{\beta^{(y)}_{D/R}}.$$

Taking the logarithm of both sides, a line is fitted using an Ordinary Least Squares (OLS) fit on the $(\log Y, \log N)$ pairs for each election for both parties or opinions (I leave the year and party notations for simplicity reasons):

$$\log(Y(N)) = \log(Y_0) + \beta \cdot \log(N),$$

where the β denotes the slope, $\log Y_0$ the the intercept of the fitted line, thus β is the exponent of the party in year y.

3 Results and discussion

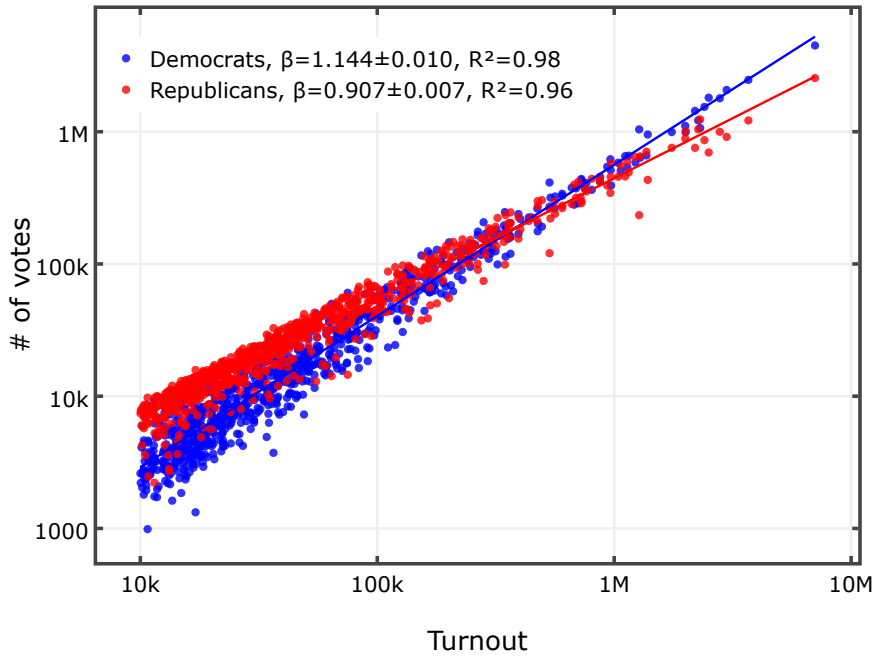


Figure 2.1. Urban scaling in the 2016 US presidential elections. Doubly logarithmic plot of votes cast for Republicans (red) and Democrats (blue) as the function of the voter turnout for the 912 largest Metropolitan and Micropolitan Statistical Areas of the US in 2016. Best OLS fit line slopes β and regression coefficients R^2 are in the insets.

Figure 2.1 shows votes for the political options as a function of voter turnout for the 912 largest Metropolitan and Micropolitan Statistical Areas representing

about 82% of the total voter population for the 2016 presidential election in the US. Figure 2.2 shows the votes as a function of voter turnout for the Remain and Leave opinions in the 2016 EU referendum for the urban electoral districts of the UK. The votes for Democrats and "Remain in the EU" scale superlinearly with exponents $\beta_D \approx 1.14$ and $\beta_{rem} \approx 1.09$, while votes for Republicans and "Leave the EU" follow sublinear scaling with $\beta_R \approx 0.92$ and $\beta_{lea} \approx 0.91$. While the elections took place in two different political situations, nevertheless they show very similar exponents.

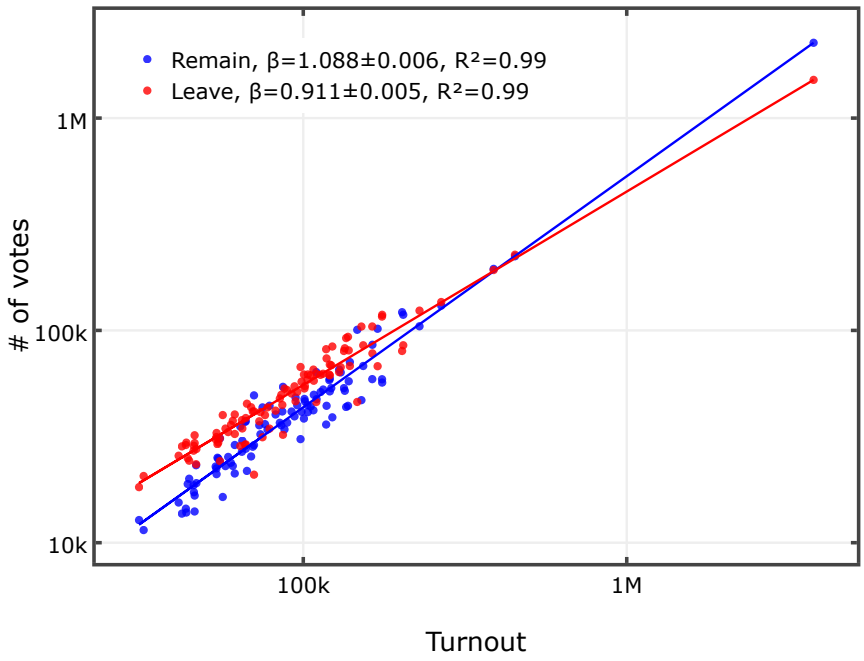


Figure 2.2. Urban scaling in the 2016 UK EU referendum. Doubly logarithmic plot of votes cast for Leave (red) and Remain (blue) as the function of the voter turnout for the the EU referendum in the UK. Best OLS fit line slopes β and regression coefficients R^2 are in the insets.

To exclude the effect of a possible scaling of the turnout with population sizes, I fitted the equation

$$Y^{(y)}(N) = Y_0^{(y)} \cdot N^{\beta_T},$$

where this time Y denotes the turnout of the election in year y, and N denotes the actual population of a city. An actual fit for the 2016 election is shown in Figure 2.3.

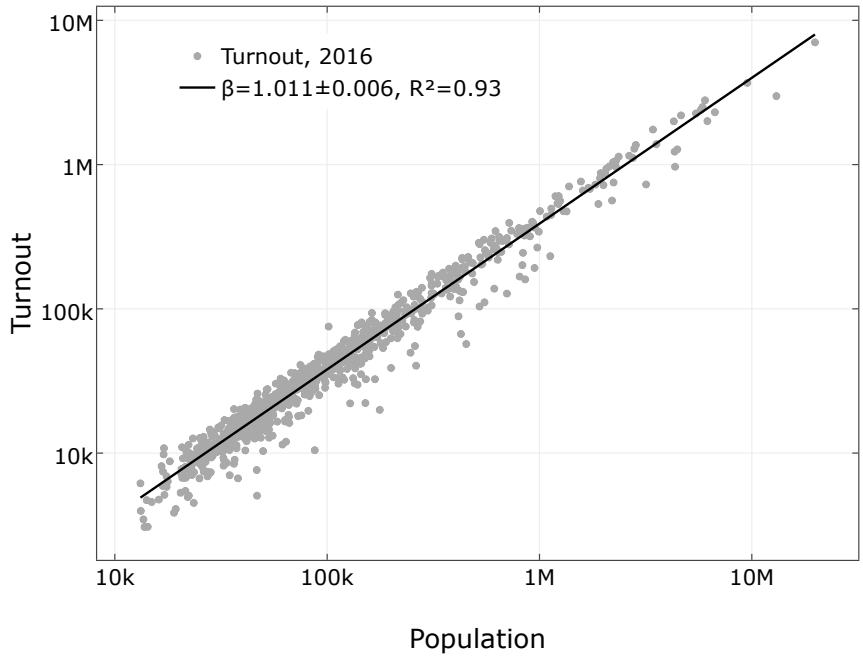


Figure 2.3. Scaling of turnout with city population in the 2016 US presidential election.

The historical exponent values are plotted in Figure 2.4.

Figure 2.5 shows the historical record of scaling exponents of the Democrats β_D and of the Republicans β_R for the 18 presidential elections in the period 1948-2016. The exponent of the Democrats has an increasing, while the exponent of the Republicans a decreasing historical trend. The Democrat and Republican curves roughly mirror each other in the whole period. The relation of the two exponents becomes apparent when the Republican exponent is plotted as a function of the Democrat exponent in Figure 2.7.

For each election and for each party I can determine the scaling exponent β and the constant Y_0 independently from the fits. Figure 2.6 shows $\log Y_0$ as a function of β, that indicates a very strong ($R^2 = 0.96$) linear relationship between the two parameters

$$\log Y_0 = -\alpha\beta + \delta, \tag{2.2}$$

for both parties and for all elections, with $\alpha = 12.111$ and $\delta = 11.396$. This linear relationship means that the parameters are not independent from each other, there-

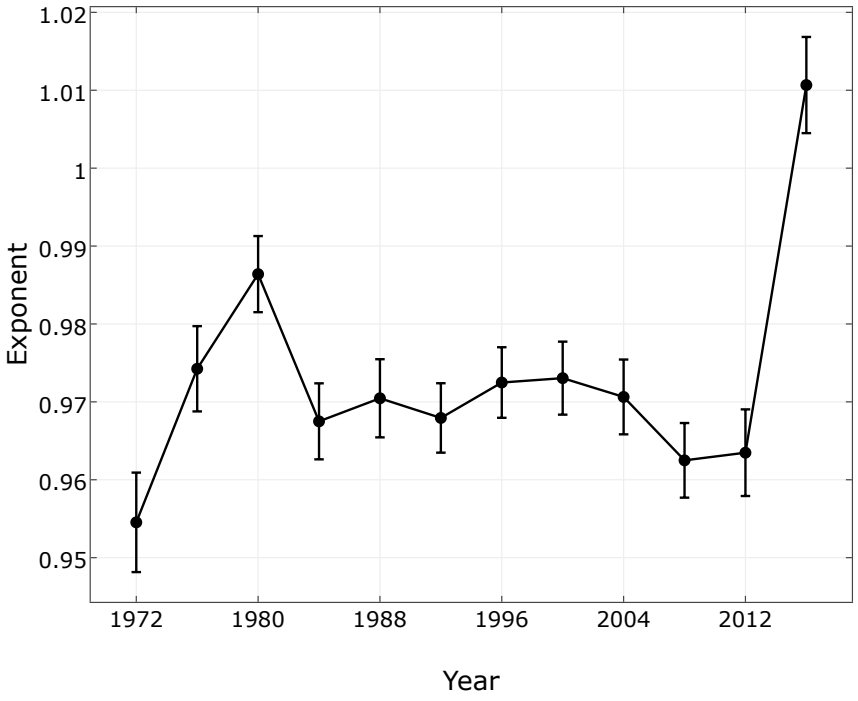


Figure 2.4. Historical scaling exponents of turnout fits in US presidential elections.

fore, given β, it is possible to determine $\log Y_0$. Thus, the number of parameters for each year's fits, which were the following four: β^D, β^R, $\log Y_0^D$ and $\log Y_0^R$ are reduced to two: β^D and β^R, from which the other two can be determined.

The above relationship (2.2) can be derived from simple analytical assumptions as well. If the intercept $\log(Y_0)$ is a function of β that changes slowly with β, and knowing that β is always close to 1, it is possible to linearly approximate $\log Y_0$ around 1:

$$\log(Y_0(\beta)) \approx \underbrace{\log(Y_0(1))}_{\delta-\alpha} + (\beta - 1) \underbrace{\left.\frac{\partial \log(Y_0(\beta))}{\partial \beta}\right|_{\beta=1}}_{-\alpha} + \cdots = -\alpha \cdot \beta + \delta$$

In the case of $\beta = 1$, it has to be true, that

$$Y_0(1) = e^{\delta-\alpha} = \langle p \rangle = p_0,$$

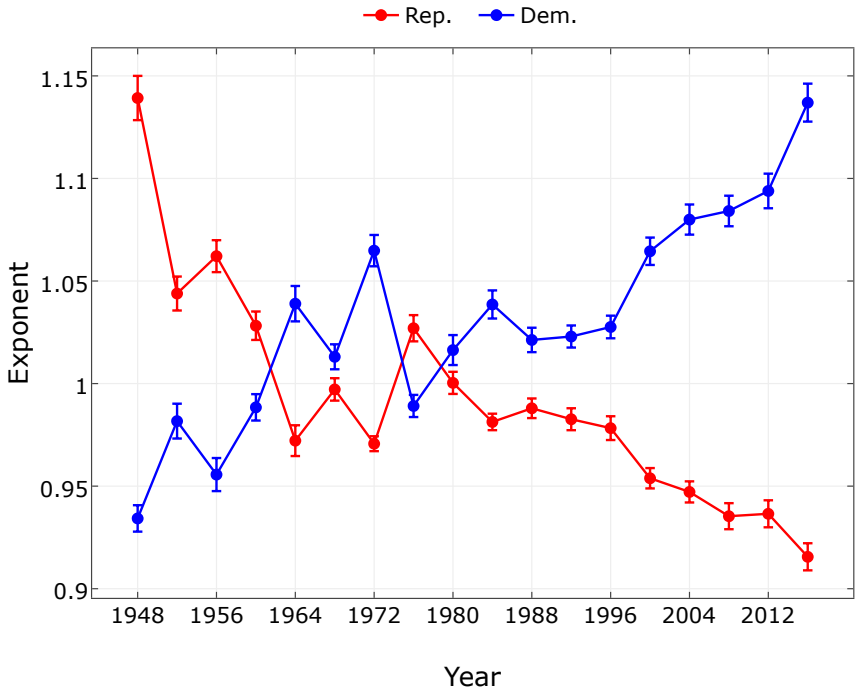


Figure 2.5. Scaling exponents for the Republicans (red) and Democrats (blue) with error bars for the 18 presidential elections of the US from 1948 to 2016.

the city-averaged voter fractions, since $\beta = 1$ would mean that every city votes as if all voters were dispersed homogeneously:

$$Y_D(N) = p_0 N.$$

Further implications can be calculated, if I let $\alpha = \log N^*$, then p_0 can also be expressed with N^*, $p_0 = e^{\delta}/e^{\alpha} = e^{\delta}/N^*$.

$$\log(Y_0(\beta)) = -\log N^* \cdot \beta + \log(p_0) + \log N^*$$

By substituting it into the original scaling relation:

$$\log Y(N) = \log(p_0 N^*) - \beta \cdot \log N^* + \beta \cdot \log N$$

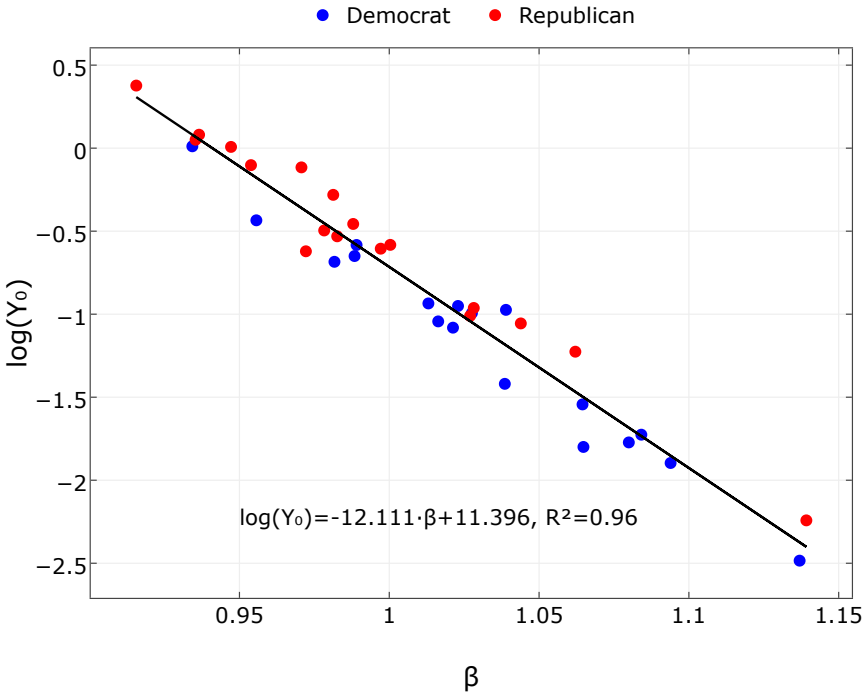


Figure 2.6. Interrelation of the exponents of urban scaling in US elections. Urban scaling exponents of Republicans as a function of the Democrats for 18 US presidential elections from 1948 to 2016 (dots) and the theoretical curve (red line) derived from probability conservation (2.5).

thus,

$$Y(N) = p_0 N^* \left(\frac{N}{N^*}\right)^{\beta} = p_0 N \left(\frac{N}{N^*}\right)^{\beta-1}$$

This implies that all fitted lines have to go through the $(N^*, p_0 N^*)$ point, because at $N = N^*$, Y equals to $p_0 N^*$ regardless of the value of β. Also note, that N^* is universal for both parties and for all elections. Thus, the scaling relations only have one parameter, the scaling exponent β. In the US historical elections, the numerical factor $e^{\delta-\alpha}$ is equal to $1/2$ within numerical error and the parameter $N^* \approx 182,000$ is the average turnout of a US city of total population $429,000$ in 2016. This is about the size of Fort Wayne IN, the 125th Metropolitan Statistical Area of the US.

Therefore, the form of the scaling relation is independent of the party and election:

$$Y = e^{\delta-\alpha} N^* \cdot \left(\frac{N}{N^*}\right)^{\beta} \approx \frac{1}{2} N^* \cdot \left(\frac{N}{N^*}\right)^{\beta}, \tag{2.3}$$

where N is the voter turnout in a city, β is the exponent of the party and $\log N^* = \alpha$.

The remarkable property of this scaling relation is that in average, at turnout $N = N^*$, the parties share the votes equally ($Y_D = Y_R = N^*/2$) independent of their exponents β_D and β_R or of the year of the election and unaffected by historic changes in population. For cities above turnout N^*, the party with higher β gets the majority of votes, while below this turnout the party with smaller β succeeds in average. While in 2016 already 125 metropolitan areas surpassed the population corresponding to this critical turnout, only 45 did so in 1948.

The observed linear relationship (2.2) and the single parameter form (2.3) of the scaling curve is predicted by the GLPLH model. Therefore, it is reasonable to assume that it can be adapted to the election process. Formally, I recover my scaling curve (2.3) from this theory by identifying the susceptible population with half of the voter turnout $N/2$ and by setting $N_0 = N^*/2$. There are two discrepancies between my scaling curve (2.3) and that of the GLPLH model. The GLPLH model is applicable for superlinear $\beta > 1$ ($Mq > 0$) values only, while in case of elections both superlinear and sublinear exponents arise, and the numerical value of $N_0 \approx 1.8 \cdot 10^5$ is nine orders of magnitude smaller for elections than for metrics covered in [97].

The main difference of elections from other urban phenomena is that the scaling curves influence each other via the competition for votes. This competition is expressed mathematically by the probability conservation

$$Y_D/N + Y_R/N = 1 \tag{2.4}$$

for the sum of the fraction of votes the parties get. Let us observe this relationship more closely!

In a given year, it holds for every city i that the number of Democrat and Republican voters is approximately equal to the turnout in the city:

$$\frac{Y_D^{(i)}}{N^{(i)}} + \frac{Y_R^{(i)}}{N^{(i)}} = 1$$

Assuming scaling from (2.3), the expected values of the Democrat and Republican voters can be substituted:

$$Y_D^{(i)} = \frac{1}{2}N^* \left(\frac{N^{(i)}}{N^*}\right)^{\beta_D} = \frac{1}{2}N^{(i)} \left(\frac{N^{(i)}}{N^*}\right)^{\beta_D - 1}$$
$$Y_R^{(i)} = \frac{1}{2}N^* \left(\frac{N^{(i)}}{N^*}\right)^{\beta_R} = \frac{1}{2}N^{(i)} \left(\frac{N^{(i)}}{N^*}\right)^{\beta_R - 1}$$

Thus,

$$\frac{1}{2}\left(\frac{N^{(i)}}{N^*}\right)^{\beta_D - 1} + \frac{1}{2}\left(\frac{N^{(i)}}{N^*}\right)^{\beta_R - 1} = 1$$
$$\left(\frac{N^{(i)}}{N^*}\right)^{\beta_D - 1} + \left(\frac{N^{(i)}}{N^*}\right)^{\beta_R - 1} = 2$$

Because the exponents β_D and β_R are close to 1, the left hand side can be approximated to the second order

$$1 + (\beta_D - 1) \cdot \log\frac{N^{(i)}}{N^*} + \frac{1}{2}(\beta_D - 1)^2 \cdot \left(\log\frac{N^{(i)}}{N^*}\right)^2 + \cdots +$$
$$1 + (\beta_R - 1) \cdot \log\frac{N^{(i)}}{N^*} + \frac{1}{2}(\beta_R - 1)^2 \cdot \left(\log\frac{N^{(i)}}{N^*}\right)^2 + \cdots = 2$$

Let us average the equation over all cities in a year:

$$(\beta_D - 1) \cdot \left\langle \log \frac{N^{(i)}}{N^*} \right\rangle + \frac{1}{2}(\beta_D - 1)^2 \cdot \left\langle \left(\log \frac{N^{(i)}}{N^*} \right)^2 \right\rangle +$$
$$(\beta_R - 1) \cdot \left\langle \log \frac{N^{(i)}}{N^*} \right\rangle + \frac{1}{2}(\beta_R - 1)^2 \cdot \left\langle \left(\log \frac{N^{(i)}}{N^*} \right)^2 \right\rangle = 0$$

In the first order, $\beta_R - 1 = -(\beta_D - 1)$. Because the term $(\beta_R - 1)^2$ is small, I only use its first order approximation, thus:

$$(\beta_D - 1) \cdot \left\langle \log \frac{N^{(i)}}{N^*} \right\rangle + \frac{1}{2}(\beta_D - 1)^2 \cdot \left\langle \left(\log \frac{N^{(i)}}{N^*} \right)^2 \right\rangle +$$
$$(\beta_R - 1) \cdot \left\langle \log \frac{N^{(i)}}{N^*} \right\rangle + \frac{1}{2}(-(\beta_D - 1))^2 \cdot \left\langle \left(\log \frac{N^{(i)}}{N^*} \right)^2 \right\rangle = 0$$

$$(\beta_D - 1) \cdot \left\langle \log \frac{N^{(i)}}{N^*} \right\rangle + (\beta_D - 1)^2 \cdot \left\langle \left(\log \frac{N^{(i)}}{N^*} \right)^2 \right\rangle +$$
$$(\beta_R - 1) \cdot \left\langle \log \frac{N^{(i)}}{N^*} \right\rangle = 0$$

This means that β_R can be approximated from β_D, which again decreases the number of parameters needed to describe a year's election scaling relationships down to one: β_D determines β_R, and from these two, the intercepts can be calculated using (2.2).

$$\beta_R - 1 = -(\beta_D - 1) \cdot +(\beta_D - 1)^2 \cdot \frac{\left\langle \left(\log \frac{N^{(i)}}{N^*} \right)^2 \right\rangle}{\left\langle \log \frac{N^{(i)}}{N^*} \right\rangle} \tag{2.5}$$

Equation (2.5) also guarantees that one of the exponents will be superlinear while the other sublinear. The numerical subsitution for the historical β_D values is shown

as a red dashed line in Figure 2.7. Since the scaling exponent of one of the parties can be determined from the other, one single parameter, the scaling exponent of one of the parties, fully determines the urban scaling curves for both parties. Accordingly, only one of the scaling exponents needs a detailed explanation. A model derived for the results of one of the parties will determine the results of the other party via probability conservation. The strategy of one of the parties will result in a superlinear exponent, which can be explained by an adaptation of the GLPLH model, while the result of the party with the sublinear exponent is just a consequence of the other party's strategy and the probability conservation law. The explanation of a strategy that results in a superlinear scaling follows later.

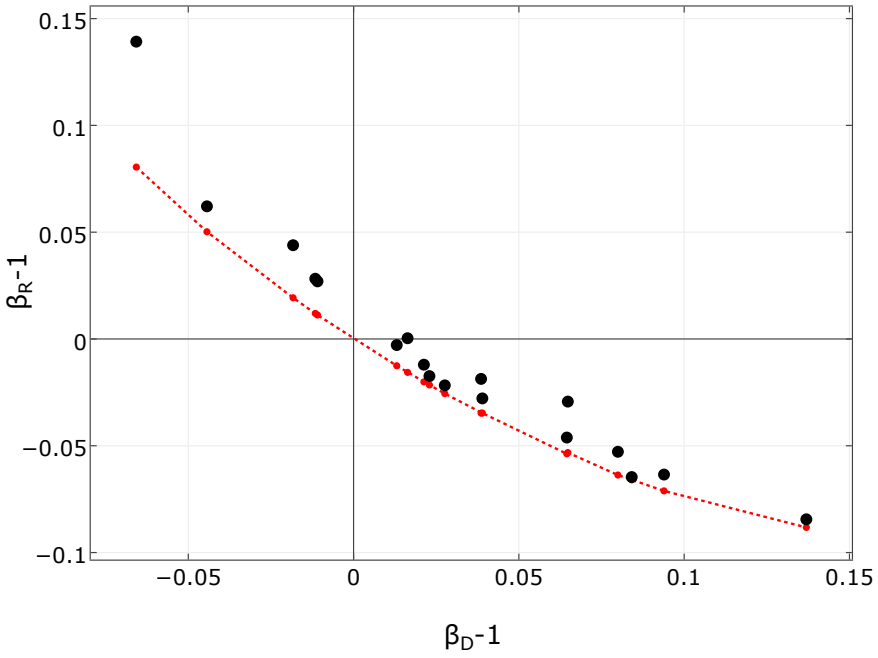


Figure 2.7. Interrelation of the parameters of urban scaling in US elections. Intercepts of the scaling relations $\log Y_0$ as a function of the scaling exponent β for Republicans (red) and Democrats (blue) for presidential elections in the period 1948-2016. Fitted line (2.2) with parameters and regression coefficient in the inset.

Next, I analyze the results in the period 2000-2016, where the Democratic party has a pronounced superlinear scaling, and the Republican party a sublinear scaling. I show that the statistical distribution of the results for Democrats is in accordance with the GLPLH model, while the distribution of the votes for Republicans deviates from it and is merely the consequence of probability conservation. I note here, that

while my fits show a superlinear scaling for the Republican party before 1960, the R^2-values of these fits are not as reliable as that of the more recent ones, and therefore, I will continue my analysis of superlinear processes and of the model only for the aforementioned years.

A Scale-Adjusted Metropolitan Indicator [101, 105–98, 137] (SAMI) is the logarithmic deviation of the value Y_i from the average scaling curve for a city with population N_i

$$\xi_i = \log Y_i - \log Y_0 - \beta \log N_i. \tag{2.6}$$

The articles [101, 105] predict that SAMIs for a given city size range are normally distributed if the investigated measure obeys the urban scaling laws. In Figure 2.9 I show the distribution of these population window-wise standardized SAMIs for both parties. Normality checks for these distributions were conducted. Based on Table 2.1, standardized SAMIs for Democratic party follow standard Gaussian distribution as confirmed by a Kolmogorov–Smirnov test in agreement with the SAMI literature. However, the same standardization procedure results in a skewed distribution for the Republicans, for whom the distribution is not normal and the GLPLH model doesn't apply. This also confirms that two parties don't have an equal role in the coupled urban scaling phenomenon.

Year	p_{Dem}	p_{Rep}
2000	0.0675	2.23e-03
2004	0.1090	3.14e-04
2008	0.1670	2.30e-04
2012	0.0746	2.05e-07
2016	0.8440	7.04e-13

Table 2.1. p-values for the Kolmogorov-Smirnov test on the distribution of the rescaled SAMIs. p_{Dem} shows the p-values for the different election years for the Democrat rescaled SAMI distributions, while p_{Rep} shows the same for the Republicans.

The GLPLH model states that the SAMI variance can be expressed with the former complexity parameter q and the number of complementary factors M as $\sigma^2_{SAMI} =$

$q^2 Mb(\log N_0 - \langle \log N \rangle)$, where $\langle \log N \rangle$ is the mean of the logarithm of city sizes. It can also be expressed with the scaling exponent

$$\sigma^2_{SAMI} = q(\beta - 1)(\log N_0 - \langle \log N \rangle). \tag{2.7}$$

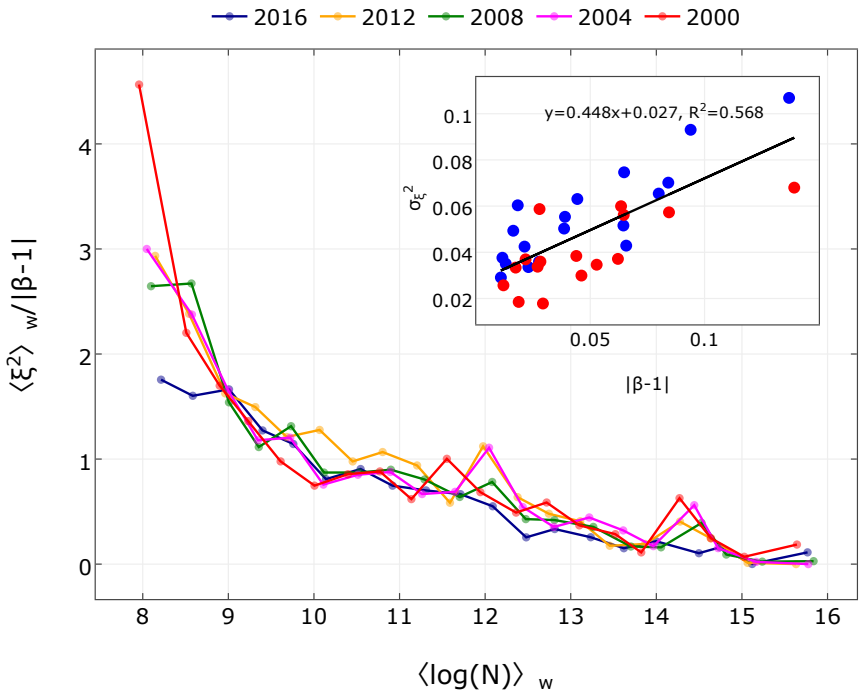


Figure 2.8. Fluctuations around the average scaling curve. City sizes are binned into 20 windows of uniform sizes on logarithmic scale. In the inset, standard deviation of SAMIs (2.6) for all metropolitan areas in my study as a function $\beta - 1$. Best fit line parameters are in the inset.

Using the inset of Figure 2.8, the general validity of this formula can be checked for both parties and for all elections in the 1948-2016 period. For the variance averaged over all metropolitan areas, though the data is noisy, it is not possible to reject the notion of proportionality with $\beta - 1$. This is also indicating that the complexity parameter q is approximately constant over several elections. From the fitted line and from the numerical value $\langle \log N \rangle = 10.55$ for the 2016 election, $q \approx 0.28$. Then, in the 2000-2016 period for the superlinearly scaling results of the Democrats, a more detailed calculation is made for ten windows of city sizes. In the main figure of Figure 2.8, the curves of $\sigma^2_{SAMI}/(\beta - 1)$ in these windows collapse onto the same curve for different elections. For low city sizes, the linear

relationship expected after the data collapse does not fit well, as shown by the inset of Figure 2.8 or even better, Figure 2.10. For greater city sizes, where statistical errors are lower, the collapse confirms that the complexity parameter q is constant. This implies that the change of the scaling exponent β_D for the Democrats in this period comes solely from the change of the number of complementary factors M.

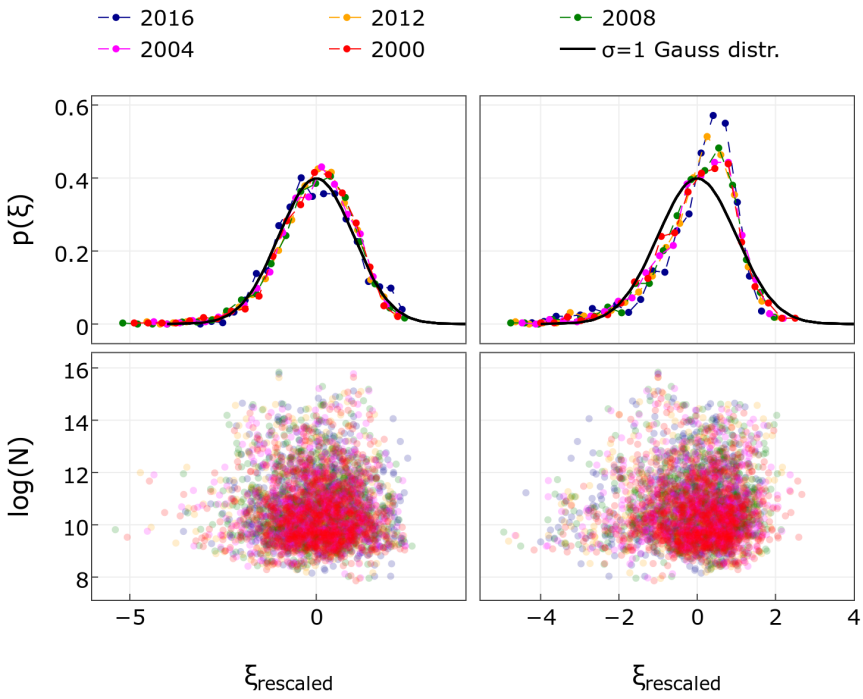


Figure 2.9. Standardized deviation of SAMIs for the last five US presidential elections. Lower panel: Scatter plot for the Democrat (left) and Republican (right) standardized deviations (horizontal axis) and logarithmic city size (vertical axis). Upper panel: Distribution of the standardized deviations. For the Democrats (left) it is a standard normal distribution (solid line). For Republicans (right) it is a skewed distribution deviating from the standard normal distribution (solid line). I used the Kolmogorov-Smirnov test to check the normality of the distributions at a significance level of 5%.

But what are the necessary complementary factors in the context of elections, that must simultaneously be present in order to vote for Democrats in the 1988-2016 period? These factors can be best understood in the terms of issue voting, where voters choose a candidate or an opinion by comparing their own viewpoints in different political issues to that of the voted one [138–131].

Because the complexity parameter q is approximately constant, the 4-6 times growth of $\beta_D - 1$ in the investigated period means that the number of issues M got

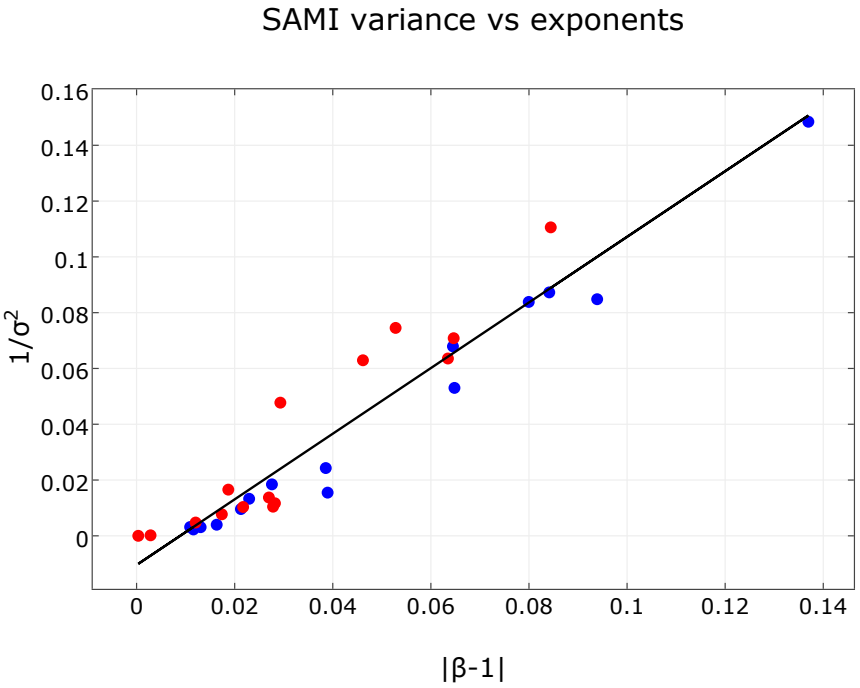


Figure 2.10. Transformed variance as a function of the exponent.

multiplicated about 4-6 times. The concrete value of M cannot be determined from the data, only the product bM that changed from about 0.09 to 0.52 with b being constant. If one could find these factors, then Republicans (or the Leave campaign) could be characterized as comprising of those voters, who don't accept at least one of those M issues that are necessary for a voter voting for the superlinearly scaling opinion or party.

Here, given the agenda of Democrats, the complementary factors might be "liberal values" in general, and the Democrat voter typically accepts all of these M values or issues simultaneously. Such values include tolerance and acceptance towards various social groups ranging from women and blacks at the beginning and middle of the 20$^{\text{th}}$ century to LGBT communities, immigrants, refugees and various other social minorities recently. If a voter is not able to accept at least one of these values or groups, then he or she will probably not vote for the Democrats. That explains why groups of the Republican and the Leave voters look so heterogeneous: they consist of groups that oppose at least one of these liberal values, and that are otherwise not held together by a common political agenda. This also explains the

recent steady increase of the Democrat exponent: as more and more M values are introduced, it increases the exponent, making bigger cities having disproportionally more Democrats than smaller ones. This result is in line with the findings of [128], who explain the same phenomenon via the increasing impact of social contagion in the more populated areas of the United States.

In this context, I can identify q as a probability that a voter – left on its own devices – rejects one of the M liberal values, and $r(N)$ is the probability that a city of size N makes a voter tolerant towards those values. Social diversity grows with the city size and voters in cities can face an increasing number of social issues and can develop tolerance towards them. This is in accordance with "the immigration paradox" phenomenon in Great Britain, where voters living near immigrants develop a tolerance towards them, while those who do not are more likely to reject them [131]. Therefore, just like other types of diversities in cities, tolerance grows like $r(N) \sim \log N/N_0$, but the number of maximal social diversity is reached at $N_0 \approx 4 \cdot 10^5$, which is smaller than the diversity $N^* \approx 1.8 \cdot 10^{14}$ observed for the more general type of diversity characterizing humans in creative aspects. Finally, there is one more consequence of this model: as the number of liberal values M seems to grow continuously, the potential voters who don't accept one of them also increases, and becomes detrimental for electoral success. This leads to the fragmentation of the political left, since a larger number of smaller parties accepting only a subset of the M values, or even "single-issue" parties can minimize the number of estranged voters and maximize the aggregated votes of all these parties.

4 Conclusion

In this chapter, I applied urban scaling theory to the number of votes cast in the Metropolitan and Micropolitan Statistical Areas in the 1948-2016 presidential elections of the US and the votes cast in the urban areas of the 2016 EU referendum in the UK. I found that out of the two voting options (Democrat/Republican, Remain/Leave), one always follows a superlinear, while the other a sublinear scaling. Using the historical dataset, I showed that instead of four parameters (two for

both scaling fits), the single exponent of the superlinearly scaling party is enough to characterize all processes across the elections and the parties. I derived the other exponent from the superlinear exponent by using the conservation of voting probabilities, and showed that it determines imposes sublinearity on the other exponent. I then analyzed the fluctuations around the scaling curve distributions and found that the distribution corresponding to the superlinear exponent is lognormal. I concluded that the two parties play different roles in urban scaling. The party with superlinear exponent drives the process, while the scaling of the party with the sublinear exponent is merely the result of probability conservation. In the context of elections I identified the elements of the GLPLH model and showed that social tolerance and diversity replace creative diversity in this context. I pointed to new political consequences of the model. I believe that the model and the calculations could further be extended to metropolitan areas in other countries or to electoral systems with multiple choices.

3

Scaling in words on Twitter

Scaling properties of language are a useful tool for understanding generative processes in texts. In this chapter scaling relations in citywise Twitter corpora coming from the Metropolitan and Micropolitan Statstical Areas of the United States are investigated. A slightly superlinear urban scaling with the city population can be observed for the total volume of the tweets and words created in a city. Then I find that a certain core vocabulary follows the scaling relationship of that of the bulk text, but most words are sensitive to city size, exhibiting a super- or a sublinear urban scaling. In both regimes, the meaning of the most superlinearly or most sublinearly scaling words is representative of their exponent. In this chapter, it is also shown that the parameters for Zipf's law and Heaps law differ on Twitter from that of other texts, and that the exponent of Zipf's law changes with city size.

The material presented in this chapter can be found in [2].

1 Introduction

The recent increase in digitally available language corpora made it possible to extend the traditional linguistic tools to a vast amount of often user-generated texts. Understanding how these corpora differ from traditional texts is crucial in developing computational methods for web search, information retrieval or machine translation [141]. The amount of these texts enables the analysis of language on a previously unprecedented scale [142–135], including the dynamics, geography and time scale of language change [145, 137], social media cursing habits [147–140] or dialectal variations [150].

Various studies have analyzed spatial variation in the text of OSN messages and its applicability to several different questions, including user localization based on the content of their posts [14, 151], empirical analysis of the geographic diffusion of novel words, phrases, trends and topics of interest [152, 144], or measuring public mood [154].

While many of the above cited studies exploit the fact that language use or social media activity varies in space, it is hard to capture the impact of the geographic environment on the used words or concepts. There is a growing literature on how the sheer size of a settlement influences the number of patents, GDP or the total road length driven by universal laws [93]. These observations led to the establishment of the theory of urban scaling [96, 88, 99, 101, 106, 98, 155–148], where scaling laws with city size have been observed in various measures such as economic productivity [137], human interactions [158], urban economic diversification [159], election data [160], building heights [161], crime concentration [162, 154] or touristic attractiveness [164].

In this chapter, the aim is to capture the effect of city size on language use via individual urban scaling laws of words. By examining the so-called scaling exponents, it is possible to connect geographical size effects to systematic variations in word use frequencies. It can be shown that the sensitivity of words to population size is also reflected in their meaning. I also investigate how social media language and city size affects the parameters of Zipf's law [165], and how the exponent of Zipf's law is different from that of the literature value [165, 157]. It is also shown

that the number of new words needed in longer texts, the Heaps law [142] exhibits a power-law form on Twitter, indicating a decelerating growth of distinct tokens with city size.

2 Methods

2.1 Twitter and census data

The data used in the present chapter was obtained from the freely available 1% sample of Twitter via the streaming API that is described in Chapter 1. Here, 456 millions of the geolocated tweets are analyzed, that have been collected between February 2012 and August 2014 from the area of the United States. A geographically indexed database is constructed of these tweets, permitting the efficient analysis of regional features [76]. Using the Hierarchical Triangular Mesh scheme for practical geographic indexing, a US county is assigned to each tweet [167, 159]. County borders are obtained from the GAdm database [169]. Counties are then aggregated into Metropolitan and Micropolitan Areas using the county to metro area crosswalk file from [135]. Population data for the MSA areas is obtained from [170].

There are many ways a user can post on Twitter. Because a large amount of the posts come from third-party apps such as Foursquare, the messages are filtered according to their URL field. Only those messages are left that have either no source URL, or their URL after the `'https://'` prefix matches one of the following SQL patterns: `'twit%'`, `'tl.gd%'` or `'path.com%'`. These are most likely text messages intended for the original use of Twitter, and where automated texts such as the phrase 'I'm at' or 'check-in' on Foursquare are left out.

For the tokenization of the Twitter messages, the toolkit published on `https://github.com/eltevo/twtoolkit` is used. Words that are less than three characters long, contain numbers or have the same consecutive character more than twice are left out. Hashtags, characters with high unicode values, usernames and web addresses [76] are also filtered.

2.2 Urban scaling

Here I investigate urban scaling relations between urban area populations and various measures of Twitter activity and the language on Twitter. As a reminder, here is the relationship from Chapter 2:

$$Y(N) = Y_0 \cdot N^{\beta}. \tag{3.1}$$

When fitting scaling relations on aggregate metrics or on the number of times a certain word appears in a metropolitan area, it is always assumed that the total number of tweets, or the total number of a certain word Y_{tot} must be conserved in the law. That means that there is only one parameter in our fit, the value of β, while the multiplication factor Y_0 determined by β and Y_{tot} as follows:

$$\sum_{i=1}^{K} Y_0 \cdot N_i^{\beta} = Y_{tot},$$

where the index i denotes different cities, the total number of cities is K, and N_i is the population of the city with index i.

For the statistical analysis, the 'Person Model' of Leitao et al. [110] is used, where this conservation is ensured by the normalization factor, and where the assumption is that out of the total number of Y_{tot} units of output that exists in the whole urban system, the probability $p(j)$ for one person j to obtain one unit of output depends only on the population N_j of the city where person j lives as

$$p(j) = \frac{N_j^{\beta-1}}{Z(\beta)},$$

where $Z(\beta)$ is the normalization constant, i.e. $Z(\beta) = \sum_{j=1}^{M} N_j^{\beta-1}$, if there are altogether M people in all of the cities. Formally, this model corresponds to a scaling relationship from (3.1), where $Y_0 = Y_{tot}/Z(\beta)$. But it can also be interpreted as urban scaling being the consequence of the scaling of word choice probabilities for a single person, which has a power-law exponent of $\beta - 1$.

To assess the validity of the scaling fits for the words, one can confirm nonlinear scaling, if the difference between the likelihoods of a model with a β_W (the scaling exponent of the total number of words) and β given by the fit is big enough. It means that the difference between the Bayesian Information Criterion (BIC) values of the two models $\Delta BIC = BIC_{\beta=1} - BIC_{\beta\neq 1}$ is sufficiently large [110]: $\Delta BIC > 6$. Otherwise, if $\Delta BIC < 0$, the linear model fits the scaling better, and between the two values, the fit is inconclusive.

2.3 Zipf's law

Here, the following form for Zipf's law is used that is proposed in [171], and that fits the probability distribution $p(f)$ of the word frequencies f apart from the very rare words:

$$p(f) = C \cdot f^{-\alpha}, \text{ if } f > f_{min},$$

where C is a constant, and α is the exponent of the power-law distribution.

The probability distribution of the frequencies is fit by using the `powerlaw` package of Python [172], that uses a Maximum Likelihood method based on the results of [112, 104, 173]. f_{min} is the frequency for which the power-law fit is the most probable with respect to the Kolmogorov-Smirnov distance [172].

A perhaps more common form of the law connects the rank of a word and its frequency:

$$f(r) = C \cdot r^{-\gamma}.$$

The first form is useful, because the fitting method of [172] can only reliably tell the exponent for the tail of a distribution. In the rank-frequency case, the interesting part of the fit would be at the first few ranks, while the most common words are in the tail of the $p(f)$ distribution.

The two formulations can be easily transformed into each other (see [171], which gives us

$$\alpha = \frac{1}{\gamma} + 1.$$

This enables us to compare our result to several others in the literature.

3 Results and discussion

3.1 Scaling of aggregate metrics

First, I checked how some aggregate metrics: the total number of users, the total number of individual words and the total number of tweets change with city size. Figures 3.1, 3.2 and 3.3 show the scaling relationship data on a log-log scale, and the result of the fitted model. In all cases, ΔBIC was greater than 6, which confirmed nonlinear scaling. The the total count of tweets and words both have a slightly superlinear exponents around 1.02. The deviation from the linear exponent may seem small, but in reality it means that for a tenfold increase in city size, the abundance of the quantity Y measured increases by 5%, which is already a significant change. The number of users scales sublinearly ($\beta = 0.95 \pm 0.01$) with the city population, though.

It has been shown in [158] that total communication activity in human interaction networks grows superlinearly with city size. This is in line with the findings that the total number of tweets and the total word count scales superlinearly. However, the exponents are not as big as that of the number of calls or call volumes in the previously mentioned article ($\beta \in [1.08, 1.14]$), which suggests that scaling exponents obtained from a mobile communication network cannot automatically be translated to a social network such as Twitter.

3.2 Individual scaling of words

For the 11732 words that had at least 10000 occurrences in the dataset, scaling relationships were fitted using the Person Model. The distribution of the fitted exponents is visible in Figure 3.5. There is a most probable exponent of approximately 1.02, which corresponds roughly to the scaling exponent of the overall word count. This is the exponent which is used as an alternative model for deciding non-

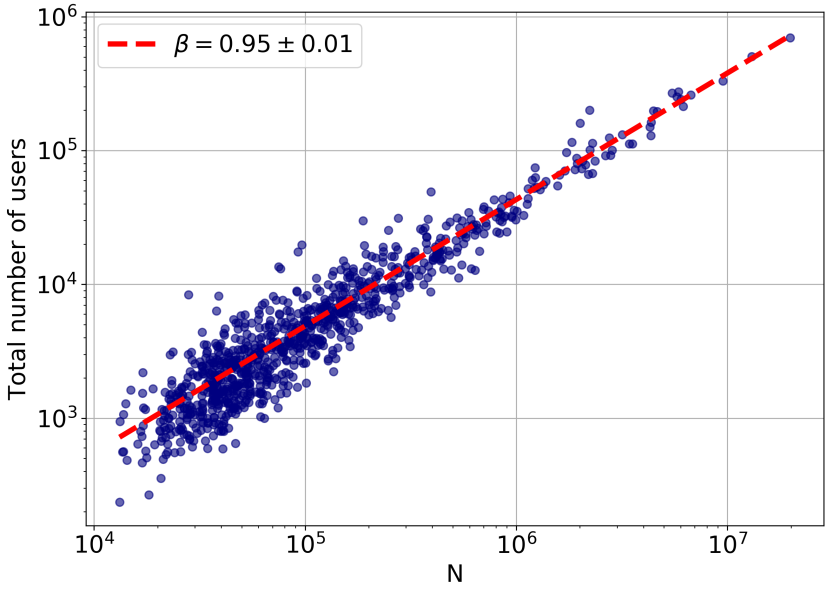


Figure 3.1. Scaling of the number of distinct users who sent a geolocated message with city population. Each point represents an MSA, the fitted line is the best MLE fit for the Person Model of [110].

linearity, because a word that has a scaling law with the same exponent as the total number of words has the same relative frequency in all urban areas. The linear and inconclusive cases calculated from ΔBIC values are located around this maximum, as shown in different colors in Figure 3.5. In this figure, linearly and nonlinearly classified fits might appear in the same exponent bin, because of the similarity in the fitted exponents, but a difference in the goodness of fit. Words with a smaller exponent, that are "sublinear" do not follow the text growth, thus, their relative frequency decreases as city size increases. Words with a greater exponent, that are "superlinear" will relatively be more prevalent in texts in bigger cities. There are slightly more words that scale sublinearly (5271, 57% of the nonlinear words) than superlinearly (4011, 43% of the nonlinear words). Three example fits from the three scaling regimes are shown in Figure 3.4.

Words falling into the "linear" scaling category were sorted according to their BIC values showing the goodness of fit for the fixed β model. The first 50 words in Table 3.1 according to this ranking are some of the most common words of the

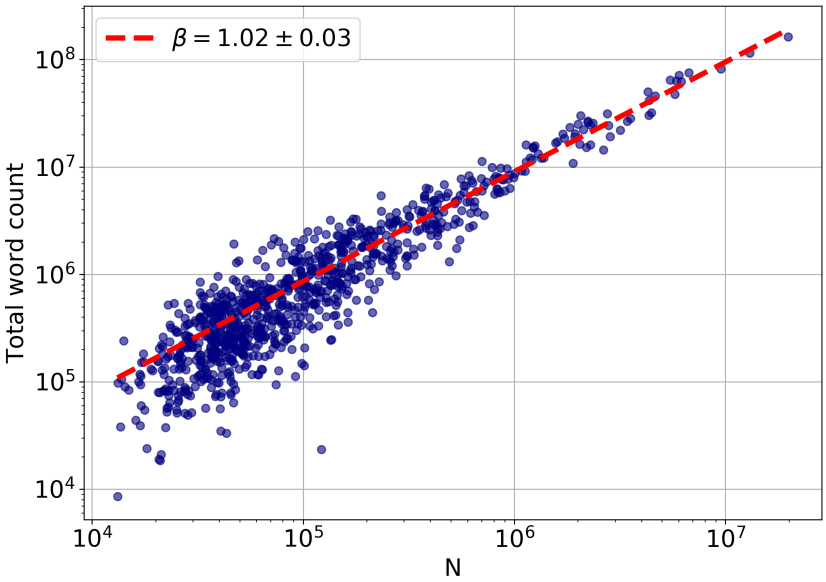


Figure 3.2. Scaling of the total number of words with city population. Each point represents an MSA, the fitted line is the best MLE fit for the Person Model of [110].

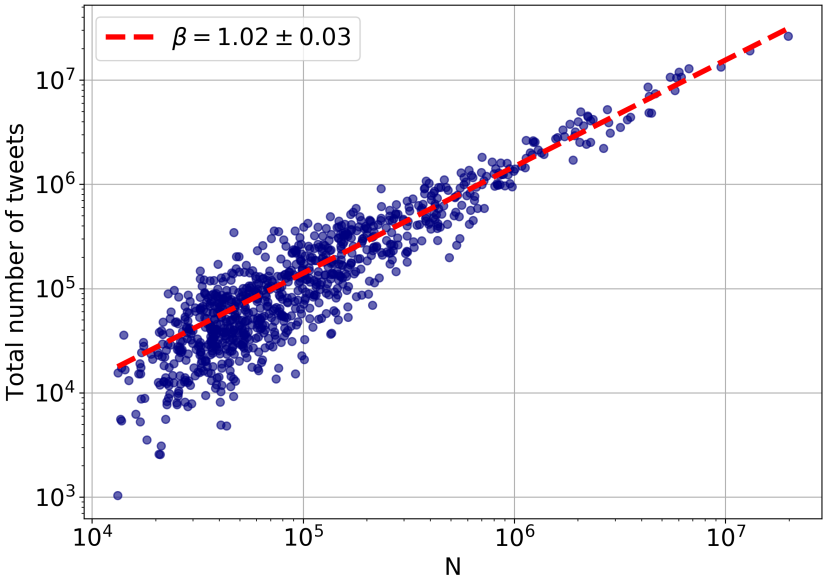
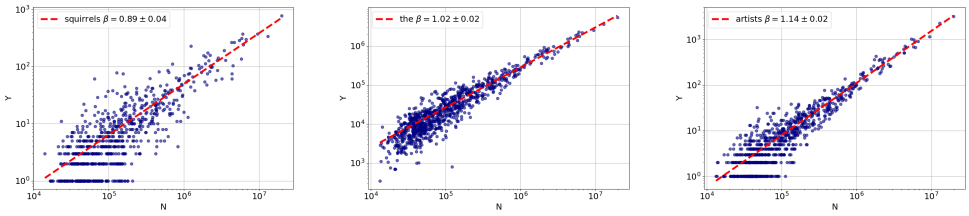


Figure 3.3. Scaling of the total number of geolocated messages with city population. Each point represents an MSA, the fitted line is the best MLE fit for the Person Model of [110].



(a) Sublinear scaling **(b)** Linear scaling **(c)** Superlinear scaling

Figure 3.4. Three scaling relationships from the sublinear (a), linear (b), and superlinear (c) scaling regimes with the MLE fits explained in the Methods section.

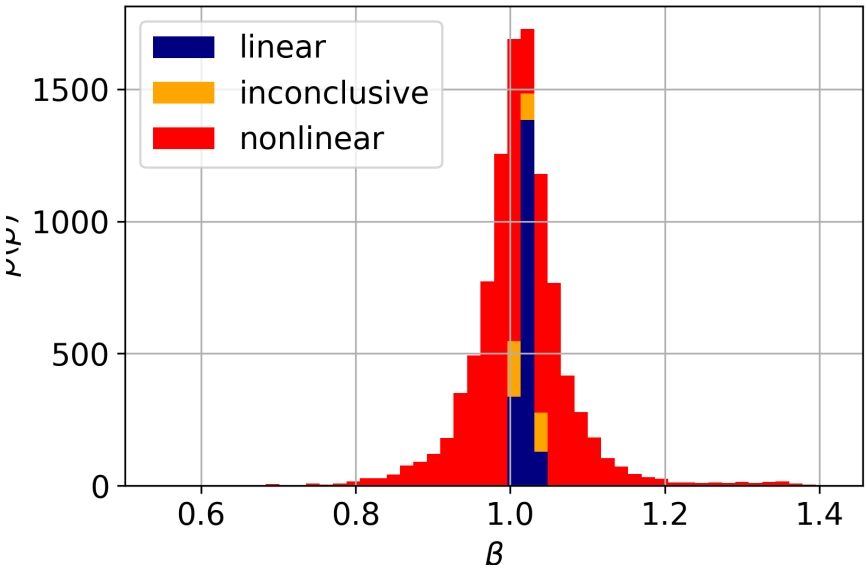


Figure 3.5. Distribution of word exponents. Statistically significant deviations from the scaling of the total number of words are marked by color codes. The peak around 1.02 marks words that have an exponent around the exponent of the total number of words. The majority of words follow a superlinear or a sublinear scaling law. Note, that there can be multiple categories in one bin according to the ΔBIC of fits.

English language, apart from some swearwords and abbreviations (e.g. lol) that are typical for Twitter language [60]. These are the words that are most homogeneously present in the text of all urban areas.

From the first 5000 words according to word rank by occurrence, the most sublinearly and superlinearly scaling words can be seen in Table 3.2. Their exponent

the you and that for this just lol like with have but get not your
was all love what are when out know good now got can about one time
day how they too shit want back need why she people right some see
going today fuck will really her

Table 3.1. The top 50 words as ranked according to the BIC values for a $\beta = 1.0207$ fixed exponent Person Model. These are the words that correspond most to the scaling of the overall word volume, thus, they are the words that appear most homogeneously in the texts of all urban areas.

differs significantly from that of the total word count, and their meaning can usually be linked to the exponent range qualitatively. The sublinearly scaling words mostly correspond to weather services reporting (flood 0.54, thunderstorm 0.61, wind 0.85), some certain slang and swearword forms (shxt 0.81, dang 0.88, damnit 0.93), outdoor-related activities (fishing 0.82, deer 0.81, truck 0.90, hunting 0.87) and certain companies (walmart 0.83). There is a longer tail in the range of superlinearly scaling words than in the sublinear regime in Figure 3.5. This tail corresponds to Spanish words (gracias 1.41, por 1.40, para 1.39 etc.), that could not be separated from the English text, since the shortness of tweets make automated language detection very noisy. Apart from the Spanish words, again some special slang or swearwords (deadass 1.52, thx 1.16, lmfao 1.17, omfg 1.16), flight-reporting (flight 1.25, delayed 1.24 etc.) and lifestyle-related words (fitness 1.15, fashion 1.15, restaurant 1.14, traffic 1.22) dominate this end of the distribution.

Thus, when compared to the slightly nonlinear scaling of total amount of words, not all words follow the growth homogeneously with this same exponent. Though a significant amount remains in the linear or inconclusive range according to the statistical model test, most words are sensitive to city size and exhibit a super- or sublinear scaling. Those that fit the linear model the best, correspond to a kind of 'core-Twitter' vocabulary, which has a lot in common with the most common words of the English language, but also shows some Twitter-specific elements. A visible group of words that are amongst the most super- or sublinearly scaling words are related to the abundance or lack of the elements of urban lifestyle (e.g. deer, fitness). Thus, the imprint of the physical environment appears in a quantifiable way in the growths of word occurrences as a function of urban populations. Swearwords and slang, that are quite prevalent in this type of corpus [147, 139], appear at both ends of the regime. It suggests that some specific forms of swearing disappear with

sublinear:

flood severe thunderstorm warning statement april lsu february bama ole unc shxt beside deer shelby kentucky ian fishing dynasty dorm freeze nigha carolina roomie walmart december january tornado gotti mountains mite wind kelsey campus exams mart roommates frat mud roads lmbo biology duke logan roommate ruzzle exam pinterest brooke bahaha slowly further mam hunting bahahaha thanking dang dwn hush softball bailey haley porch rec gates yuu november memphis marshall haven storms ncaa cody renee tanning oomf heck paige nosey casino southern muh bre lab tub truck cowboy jeep seth messy lawd layin tourney trashy puke library gah lake tweeps rae semester wreck johns bonfire studying until quit state gotcha anatomy prolly knw eagle wrk lifting flag lastnight courtney awhile tweetin bend ann abby march douche snuggle fog bracket hannah bedtime golf sittin gosh lynn whiskey nerves rain road town fixing hut whatcha drinkin driveway damnit country moore riley lyin duck

superlinear:

hoy gracias por para feliz con cuando que siempre verdad algo donde amor ver tiempo mejor semana estas alguien bien jajaja mas del todo jajajaja vez tus ama tengo vamos porque buenos eres linda muy quiero puedo hola las mucho nada sabes mañana amo soy les tambien vas dormir buenas amigo hay madrid mis bueno gusta brunch mal jaja uno flight familia dos cara delayed landed dice casa amigos loco grande papi fin traffic tix com lounge puerto heights brazil rico deja gate madre solo pls luis plane event international bon bella oscar sin mil damm ily mon studio maria carlos lmfao italian das film thx omw peep era salon omfg van jose london sushi blocks security vip mah ilysm hookah fitness cos ariana fashion via park jenny performing pronto artists stadium kanye restaurant awk melissa market danny ale booked leo inspired connect rft fab culture artist demi blasting design

Table 3.2. The most sublinearly ($0.54 < \beta < 0.93$) or superlinearly ($1.13 < \beta < 1.41$) scaling words out of the 5000 most frequent words with small bootstrapped error $\Delta\beta < 0.1$. Sublinear words are sorted in an ascending, superlinear words in a descending order with respect to β.

urbanization, but the share of overall swearing on Twitter grows with city size. The peak consisting of Spanish words at the superlinear end of the exponent distribution marks the stronger presence of the biggest non-English speaking ethnicity in bigger urban areas. This is confirmed by fitting the scaling relationship to the Hispanic or Latino population [174] of the MSA areas ($\beta = 1.31 \pm 0.14$), which despite the large error, is very superlinear.

3.3 Zipf's law on Twitter

Figure 3.6 shows the distribution of word counts in the overall corpus. The power-law fit gave a minimum count $x_{min} = 13$, and an exponent $\alpha = 1.682 \pm 0.001$. To check whether this law depends on city size, the same distribution was fitted for the individual cities, and according to Figure 3.7, the exponent gradually decreases with city size, that is, it decreases with the length of the text.

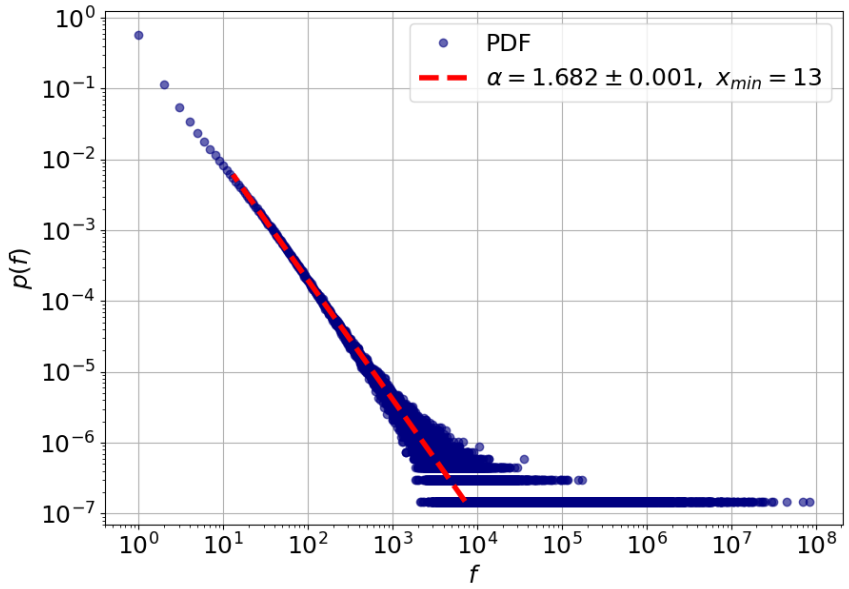


Figure 3.6. Probability distribution of word frequencies in the overall corpus and power-law fitted by the `powerlaw` package.

That the relative frequency of some words changes with city size means that the frequency of words versus their rank, Zipf's law, can vary from metropolitan area to

metropolitan area. Thus, the exponent of Zipf's law depends on city size, namely that the exponent decreases as text size increases. It means that with the growth of a city, rarer words tend to appear in greater numbers. The values obtained for the Zipf exponent are in line with the theoretical bounds 1.6-2.4 of [175]. In the communication efficiency framework [175, 167], decreasing β can be understood as decreased communication efficiency due to the increased number of different tokens, that requires more effort in the process of understanding from the reader. Using more specific words can also be a result of the 140 character limit, that was the maximum length of a tweet at the time of the data collection, and it may be a similar effect to that of texting [177]. This suggests that the carrying medium has a huge impact on the exact values of the parameters of linguistic laws.

The Zipf exponent measured in the overall corpus is also much lower than the $\beta = 2$ from the original law [165]. The second power-law regime cannot be observed either, as suggested by [178] and [171]. Because most observations so far hold only for books or corpora that contain longer texts than tweets, our results suggest that the nature of communication, in our case Twitter itself affects the parameters of linguistic laws.

3.4 Vocabulary size change

Figure 3.8 shows the vocabulary size as a function of the metropolitan area population, and the power-law fit. It shows that in contrary to the previous aggregate metrics, the vocabulary size grows very sublinearly ($\beta = 0.68$) with the city size. This relationship can also be translated to the dependency on the total word count, which would give a $\beta = 0.68/1.02 = 0.67$, another sublinear scaling.

The decrease in β for bigger cities (or bigger Twitter corpora) suggesting a decreasing number of words with lower frequencies is thus confirmed. There is evidence, that as languages grow, there is a decreasing marginal need for new words [179]. In this sense, the decelerated extension of the vocabulary in bigger cities can also be regarded as language growth.

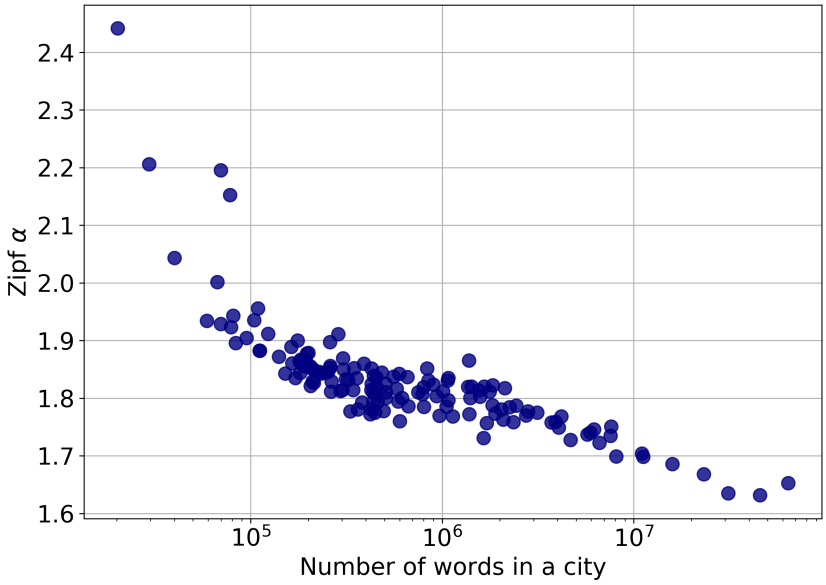


Figure 3.7. Dependency of the Zipf exponent on city population. The exponent decreases as the number of words in a city grows.

4 Conclusion

In this chapter, I investigated the scaling relations in citywise Twitter corpora coming from the Metropolitan and Micropolitan Statstical Areas of the United States. A slightly superlinear scaling could be observed decreasing with the city population for the total volume of the tweets and words created in a city. When observing the scaling of individual words, a certain core vocabulary following the scaling relationship of that of the bulk text has been found, but most words are sensitive to city size, and their frequencies either increase at a higher or a lower rate with city size than that of the total word volume. At both ends of the spectrum, the meaning of the most superlinearly or most sublinearly scaling words is representative of their exponent. I also examined the increase in the number of words with city size, which has an exponent in the sublinear range. This shows that there is a decreasing amount of new words introduced in larger Twitter corpora.

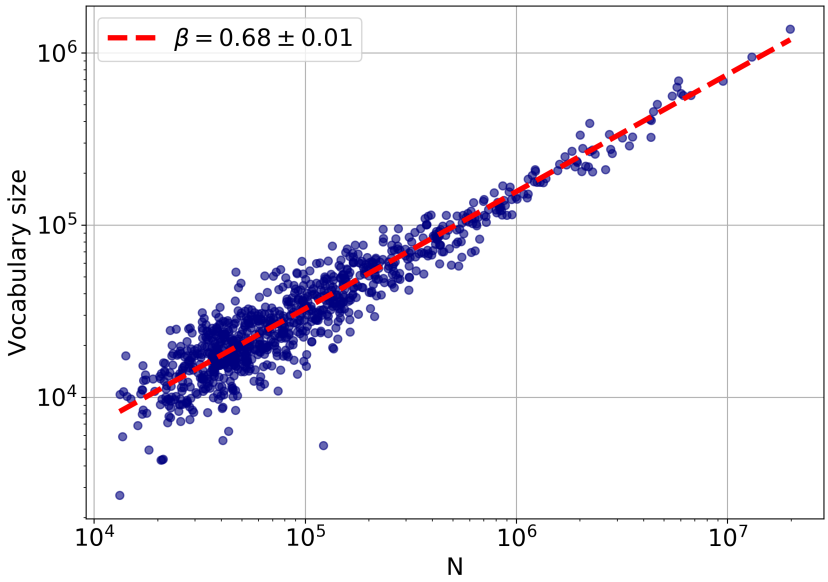


Figure 3.8. Scaling of the total number of distinct words with city population. Each point represents an MSA, the fitted line is the best MLE fit for the Person Model of [110].

4

Unemployment rates from Twitter daily rhythms

By modeling macro-economical indicators using digital traces of human activities on mobile or social networks, it is possible to provide important insights to processes previously assessed via paper-based surveys, polls or infrequently updated official records. For this chapter, aggregated workday activity timelines of US counties have been collected from the normalized number of messages sent in each hour on the online social network Twitter. I show that county employment and unemployment statistics are encoded in the daily rhythm of people by decomposing the activity timelines into a linear combination of two dominant patterns. The mixing ratio of these patterns defines a measure for each county, that correlates significantly with employment (0.46 ± 0.02) and unemployment rates (-0.34 ± 0.02). Thus, the two dominant activity patterns can be linked to rhythms signaling the presence or lack of regular working hours of individuals. The analysis could provide policymakers a better insight into the processes governing employment, where problems could not only be identified based on the number of officially registered unemployed people, but also on the basis of the digital footprints people leave on different platforms.

The material presented in this chapter appeared in [3].

1 Introduction

Several aspects on the possible usage of mobile phone records and social media status updates in the estimation of official data, such as census, demographic or land use records have been discussed in recent papers. A promising approach is the analysis of the diurnal rhythm of humans. Due to the 24 hour periodicity of the Earth's rotation, we are biologically bound to show daily periodic behavior both at the individual and at the aggregate level. This periodic cycle is governed mainly by internal biochemical processes [180–174], but the impact of external factors and the environment also leaves its imprint on these daily patterns [184, 176].

As Säramaki and Moro point out in their paper [186], an interesting application is to consider the geospatial aspects of the aggregate level of daily rhythms, as it can provide insight into several different phenomena ranging from the actual land use patterns in a city [187–188] and on a campus [189], to the tracking of anomalous events [197, 189], or the estimation of population size [199], mobility patterns [200], poverty [201] or crime rates [202] in a certain area.

Because regardless of the phenomenon, these aggregate spatio-temporal patterns always consist of the superposition of the daily rhythms of individuals, it is worth investigating how the main features of the aggregate level form from superposition. If individuals could be clustered into more or less homogeneously behaving groups based on their daily activity patterns or rhythms [203], then the aggregate pattern can be understood as the combination of the group patterns, and the group that has more individuals dominates the aggregate daily rhythm. The groups of individuals can form along many demographic and/or socioeconomic factors, of which being employed and going to and from work at regular hours is the most determining one with respect to the daily activity patterns. Thus, decomposing the groups from the aggregate patterns in different geographical regions might give insight into the estimation of employment statistics in that region.

Nowcasting or estimating unemployment rates using the digital traces of search engines has already been in the focus of several papers [204–197]. It has already been shown, that daily activity patterns of individuals can be linked to the regularity of their working hours [39]. The loss of a job has severe psychological consequences

[207], therefore, a mass layoff can be detected based on the observation of the mobile call patterns of individuals [208]. Such a mass layoff also appears in unemployment statistics, though much later than it is predictable from the mobile call data. In [209], there is a strong evidence that aggregated daily activities of certain time intervals of geographical regions can be indicative of unemployment rates.

In this chapter, 63 million geolocated messages are obtained from the publicly available stream of the social network Twitter from the area of the United States sent between January and October 2014. Monday to Friday relative tweeting activity is aggregated for each hour in each US county to form an average workday activity pattern. All users are included into the aggregation process, thus, we did not separate the users based on their age or employment status. It is then assumed that the daily activity patterns form a roughly linear subspace of the 24-hour "timespace". By finding this linear subspace, that is, by finding the line on which the county patterns lie, it is possible to give a measure that is linked to the ratio of two groups of people tweeting in a county. This measure correlates significantly with county employment and unemployment rates, and the average patterns corresponding to the two groups can be linked to lifestyles connected to regular working hours or the lack of them. Thus, it is possible to give a framework for decomposing the digital activity patterns of geographical regions and linking the decomposition to employment and unemployment rates.

2 Methods

2.1 Twitter dataset

In this analysis, I use the 1% datastream that is freely provided by Twitter through their Application Program Interface described in Chapter 1. In this study, I focus on the part of the data stream that has geolocation information. The dataset includes a total of 63 million tweets from the contiguous United States collected between January 2014 and October 2014. Using the Hierarchical Triangular Mesh scheme for practical geographic indexing [167, 159], a US county was assigned to each tweet. County borders are obtained from the GAdm database [169].

2.2 Demographic datasets

For the population-weighed linear model of the next section, county-level population statistics was obtained from the US 2010 Census [210]. Unemployment and labor force data were downloaded for the time window of the Twitter dataset from the Local Area Unemployment Statistics page of the Bureau of Labor Statistics [211]. An average was taken for the months ranging from January 2014 to October 2014 for each county.

Though unemployment levels are defined as the number of unemployed per total labor force in a county, in this chapter, the share of employed is defined as the number of employed people divided by the whole population of a county. This measure fits the model for the daily rhythm better as discussed in the Results section.

2.3 Daily activity patterns

Let's define a daily activity pattern with an hourly resolution for each county, that are enumerated by $k = 1 \dots M$. All tweets originating from a given county from the period between January 2014 and October 2014 are counted in each hour (the hour range goes from $i = 0, 1, \dots, 23$) on workdays, that is from Monday to Friday, after correcting for timezone and daylight saving time in each county. Because of the differing population and Twitter penetration rates (share of people using Twitter) in each county, the number of tweets n_i is normalized by the total tweet counts. Thus, each county (k) is represented by a 24-dimensional vector ($\mathbf{y}^{(k)}$), where the elements of $\mathbf{y}^{(k)}$ are the normalized hourly activity ratios:

$$y_i^{(k)} = \frac{n_i}{\sum_{i=0}^{23} n_i},$$

and obviously,

$$\sum_{i=0}^{23} y_i^{(k)} = 1 \quad \forall k = 1 \dots M.$$

To improve the quality of the dataset, only those counties are considered in which the overall tweet count during the ten month exceeded the threshold of 1800. Thus, 1884 counties are left for the analysis.

2.4 Linear model

The tweeting pattern of a county is supposed to be represented by the linear combination of only two universal patterns (**A** and **B**, where boldface denotes vector quantities) that are mixed for each county k with a proportion of $\alpha^{(k)}$, and $1 - \alpha^{(k)}$, respectively. Thus, the two universal patterns that compose the pattern of a county are identified as corresponding to two differently behaving population groups, whose aggregate tweeting patterns form **A** and **B**. There is no further restriction on these $\alpha^{(k)}$ values, they can be any arbitrary real numbers.

Then the predicted activity $x_i^{(k)}$ of a county k in hour i would be the following linear combination:

$$x_i^{(k)} = \alpha^{(k)} \cdot A_i + (1 - \alpha^{(k)}) \cdot B_i = \alpha^{(k)}(A_i - B_i) + B_i. \tag{4.1}$$

Let us denote the weight of each county by $w^{(k)}$, which is proportional to its population $p^{(k)}$, such that $w^{(k)} = p^{(k)} / \sum_{k=1}^{M} p^{(k)}$. The squared error of the model is then defined as

$$E = \sum_{i,k} w^{(k)} \left(y_i^{(k)} - \underbrace{\left(\alpha^{(k)}(A_i - B_i) + B_i\right)}_{x_i^{(k)}} \right)^2 .$$

The aim is to minimize this error with subject to the two conditions $\sum_i A_i = 1, \sum_i B_i = 1$. This leads to the following expression to minimize with Lagrange multipliers λ_a and λ_b:

$$E + \lambda_a \left(\sum_i A_i - 1 \right) + \lambda_b \left(\sum_i B_i - 1 \right) = \min. \tag{4.2}$$

The derivatives yield the following linear equation system:

$$\frac{\partial}{\partial A_j}: \qquad \sum_k 2w^{(k)} \left(y_j^{(k)} - \alpha^{(k)}(A_j - B_j) - B_j\right)\left(-\alpha^{(k)}\right) + \lambda_a = 0 \tag{4.3}$$

$$\frac{\partial}{\partial B_j}: \qquad \sum_k 2w^{(k)} \left(y_j^{(k)} - \alpha^{(k)}(A_j - B_j) - B_j\right)\left(-(1 - \alpha^{(k)})\right) + \lambda_b = 0 \tag{4.4}$$

$$\frac{\partial}{\partial \alpha^{(m)}}: \qquad \sum_i 2w^{(m)} \left(y_i^{(m)} - \alpha^{(m)}(A_i - B_i) - B_i\right)\left(-(A_i - B_i)\right) = 0 \tag{4.5}$$

Summing Eq 4.3 and Eq 4.4 for j yield 0 for the Lagrange multipliers λ_a and λ_b. Thus, the problem reduces to minimizing E, which actually measures the sum of squared distances from the line parametrized by $\mathbf{A} - \mathbf{B}$, $\mathbf{B}$ and $\alpha^{(k)}$ for a county k.

Since

$$\sum_j [(4.3) + (4.4)] \cdot (A_j - B_j) = \sum_m (4.5), \tag{4.6}$$

the equation system is not linearly independent. Thus, exact values for A_j, B_j and $\alpha^{(k)}$ can't be obtained, they will be dependent on each other.

Expressing $\alpha^{(k)}$ from the equation system yields:

$$\alpha^{(m)} = \frac{\sum_i \left(y_i^{(m)} - B_i\right)(A_i - B_i)}{\sum_i (A_i - B_i)^2} = \frac{(\mathbf{y}^{(m)} - \mathbf{B})(\mathbf{A} - \mathbf{B})}{(\mathbf{A} - \mathbf{B})^2}. \tag{4.7}$$

The line from which the summed distance of the datapoints is minimal is the line whose direction is parallel to the eigenvector ($\mathbf{m}$) corresponding to the largest eigenvalue of the covariance matrix $\mathbf{C}$, where

$$C_{ij} = \langle y_i y_j \rangle - \langle y_i \rangle \langle y_j \rangle, \tag{4.8}$$

if $\langle\rangle$ denotes the weighted mean ($\sum_k w^{(k)} = 1$, $w^{(k)} \geq 0 \, \forall k = 1 \ldots M$)

$$\langle y_j \rangle = \sum_k w^{(k)} y_j^{(k)}. \tag{4.9}$$

By substituting the expression for $\alpha^{(k)}$ into Eq 4.3)+Eq 4.4, and by averaging over k, the point $\langle \mathbf{y} \rangle$ should fit onto the line.

Thus, a valid solution of the error minimization problem is, if

$$\sigma(\alpha) = \sqrt{\frac{\sum_{k=1}^{M} \left(\alpha^{(k)}\right)^2}{M}},$$

$$\mathbf{A} = \langle \mathbf{y} \rangle + 2 \cdot \mathbf{m} \cdot \sigma(\alpha), \tag{4.10}$$

$$\mathbf{B} = \langle \mathbf{y} \rangle - 2 \cdot \mathbf{m} \cdot \sigma(\alpha), \tag{4.11}$$

and calulate $\alpha^{(k)}$ values according to Eq 4.7.

In other words, the minimum occurs if $\mathbf{A} - \mathbf{B}$ is parallel to the eigenvector $\mathbf{m}$ corresponding to the biggest eigenvalue of the weighed covariance matrix $\mathbf{C}$, and that $\mathbf{B}$ can be chosen as the average of $\mathbf{y}^{(k)}$s. Here, an element of the covariance matrix $\mathbf{C}$ is

$$C_{ij} = \langle y_i y_j \rangle - \langle y_i \rangle \langle y_j \rangle , \tag{4.12}$$

where

$$\langle y_j \rangle = \sum_k w^{(k)} y_j^{(k)}. \tag{4.13}$$

Consider now a linear representation of the data with a coordinate system where the mean $\langle \mathbf{y} \rangle$ sets the origin and $\mathbf{m}$ is the direction of the line. $\alpha^{(k)}$ values are calculated for each county by projecting $y^{(k)}$ onto this line. A positive $\alpha^{(k)}$ means a county, where the majority of people are active on Twitter in correspondence with the daily rhythm dictated by $\mathbf{m}$, accordingly, negative $\alpha^{(k)}$ is in connection with an opposite pattern.

Because the linear equation system derived from the minimization of the squared error is linearly dependent, the scale on the line is not set, as $\mathbf{A} - \mathbf{B}$ is only determined up to an arbitrary scaling factor. Thus, the $\alpha^{(k)}$ values are also determined only up to a scaling factor. Let us now choose $\mathbf{A}$ and $\mathbf{B}$ to be two standard deviations of $\alpha^{(k)}$-s away from the origin $\langle \mathbf{y} \rangle$ in the two directions of the new linear

coordinate system:

$$\sigma(\alpha) = \sqrt{\frac{\sum_{k=1}^{M} \left(\alpha^{(k)}\right)^2}{M}},$$

$$\mathbf{A} = \langle \mathbf{y} \rangle + 2 \cdot \mathbf{m} \cdot \sigma(\alpha), \tag{4.14}$$

$$\mathbf{B} = \langle \mathbf{y} \rangle - 2 \cdot \mathbf{m} \cdot \sigma(\alpha). \tag{4.15}$$

A and **B** are both normalized to 1, where in the 2-dimensional case their components represent the selected two hours, while in the 24 dimensional case they represent the 24 hours of the day.

3 Results and discussion

In this section, I present the description and the discussion of the main results of this chapter. First, the correlation between the activities of individual hours and employment and unemployment rates is investigated, and two dimensions with which employment and unemployment levels have maximum or minimum correlations are chosen. It is then evaluated, to what extent the linear model is a valid description of the data for these most separating dimensions (2) and then for all possible dimensions (24) of the dataset. Second, there is a discussion on how the linear models in 2 and 24 dimensions separate the two population groups with the two distinct activity patterns, and a possible interpretation of these patterns is given. Third, the two groups are connected with real-world indicators like share of employed in a county, and the plausibility of the correspondence of the daily patterns of the two separate groups to employment status is discussed.

First, population-weighted Pearson correlations are evaluated for each hour i between $y_i^{(k)}$ activities for the 1884 counties (from which the number of messages is adequate) and employment and unemployment levels. The errors of these correlations are calculated by bootstrapping the sample for $n = 1000$ times, the results with errorbars are shown in Figure 4.1. While unemployment levels are defined in the traditional way of the Bureau of Labor Statistics, here, the share of employed is defined slightly differently, normalizing the number of employed by the entire

population of a county. This definition matches the notion of population share of "active" people regarding regular working hours better.

The hours between 6am and 8pm show a significantly positive correlation with employment, and a negative one with unemployment. During the night, between 9pm and 5am, the correlation is reversed. With respect to employment, the correlation peaks at 12pm with 0.43 ± 0.02 and reaches its lowest value at 1am with -0.39 ± 0.03. The location of the maximum and minimum of correlation with unemployment are shifted slightly to 12pm and 12am, though exactly with opposite signs (0.30 ± 0.02 for 12am and -0.38 ± 0.02 for 12pm). The signs of the correlations and the hours of their extreme values indicate that increased daytime activity is associated with higher employment levels, and higher than average nighttime activity corresponds to higher unemployment.

Figure 4.1. Population-weighted Pearson correlation of employment and unemployment levels with hourly activities. Errorbars are calculated using bootstrapping $n = 1000$ times. The hours between 6am and 8pm correlate significantly positively with employment and negatively with unemployment. This relationship turns out to be exactly the opposite during the night. Regarding employment, the most distinguishing hours are 1am (most negative correlation) and 12pm (most positive correlation).

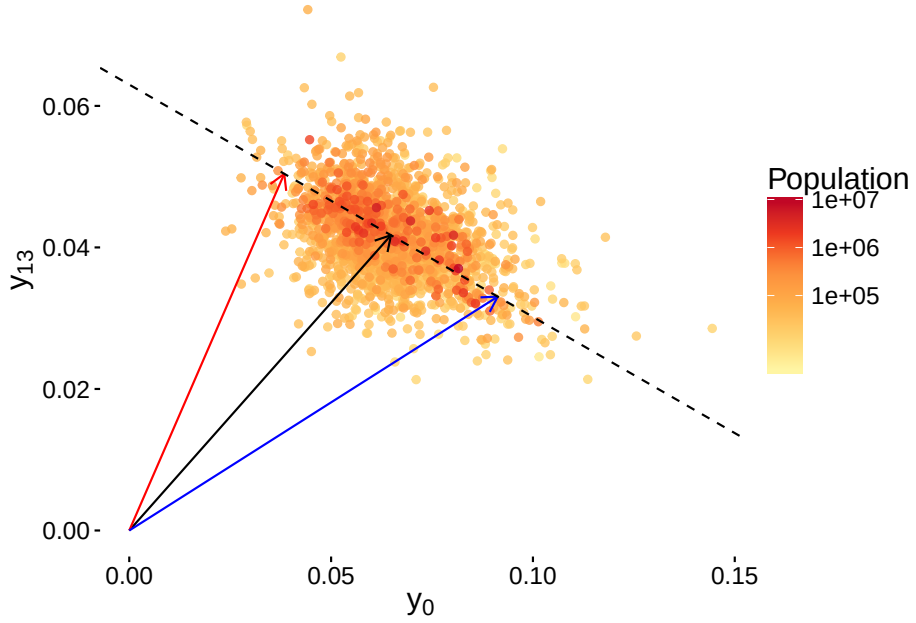


Figure 4.2. Activity of counties in the space of 12am and 1pm. Each dot represents a county, and the horizontal axis measures the relative tweeting activity between 12am and 1am in that county, while the vertical axis represents the relative tweeting activity between 1pm and 2pm in that county. As these two measures are correlated, a linear transformation cound combine them into a single coordinate. The new coordinate axis is represented by the dashed line. The black arrow points to the average of the measures along the original axes. The blue and the red arrows are possible choices for **A** and **B** vectors, see the Linear model part in the Methods section.

To check the linearity of the model described in the Methods section, the coordinate system of the hours having the extreme correlation values with employment levels is chosen at first. Figure 4.2 shows the 12am (y_0) and 1pm (y_{13}) activities of the filtered counties with the dashed line corresponding to the direction of the first eigenvector of the covariance matrix, now calculated only from these two dimensions. Eigenvalues are normalized by their sum. The first eigenvalue of the covariance matrix carries 0.99 share from all the variance in the data. Thus, linearity in this two-dimensional subspace of the whole 24-hour activity space is a good assumption.

Then the validity of the linear model in all 24 dimensions presented in Eq 4.1 is assessed. In Figure 4.3a eigenvalues of the covariance matrix **C** are plotted, again, normalized by the sum of all eigenvalues. Only the first four eigenvalues correspond

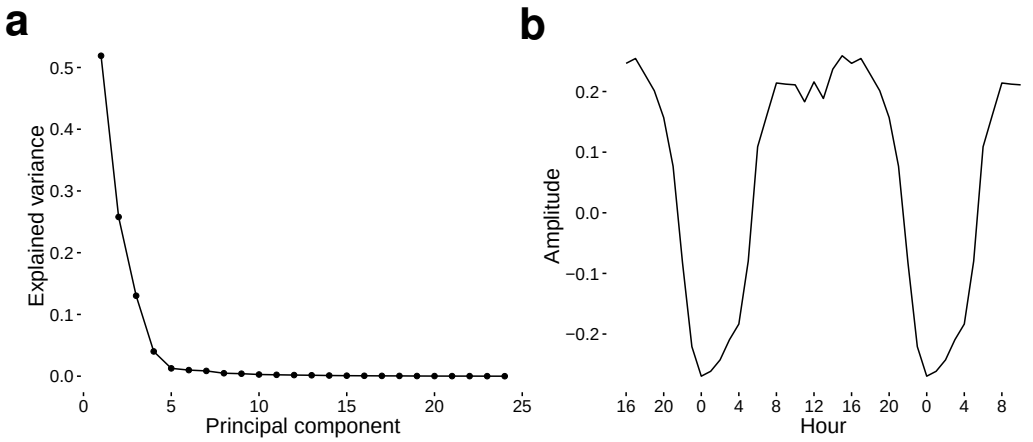


Figure 4.3. The result of the principal component analysis of the population-weighted covariance matrix. **a** Proportion of explained variance for the principal components of the covariance matrix. Only the first four components carry a share of variance significantly greater than zero. **b** Principal component corresponding to the largest principal value. The amplitude of vector components is plotted, each vector coordinate corresponds to an hour ranging from 0 to 23. The vector components are positive from 5am to 8pm, and negative otherwise.

to a variance significantly greater than 0, and the first principal component stands out with a proportion of 0.52, whereas the other three significant components carry 0.25, 0.13 and 0.04 share of the variance. Thus, the dataset is mostly linear even in the 24-dimensional space, and the representation with Eq 4.1 remains plausible.

In the 2-dimensional case, the dashed line of Figure 4.2 marks the direction of the first principal vector. The difference between the two vectors **A** (red) and **B** (blue) representing the two universal patterns (see Methods on p. 66) is parallel to this component, let us denote it by **m**. It can be seen in Figure 4.2 that the 2-dimensional **A** pattern is marked by an increased activity at 1pm, and a decreased activity at 12am, while pattern **B** is characterized by exactly the inverse relationship.

The principal component corresponding to the largest principal value in the 24-dimensional case can be seen in Figure 4.3. As the coordinates represent the hours, it can be seen from Figure 4.3 that **m** is positive from 5am until 8pm, and negative otherwise. Thus, the positive elements of **m** select mainly those hours during which people are awake, and the negative elements correspond to the sleeping hours.

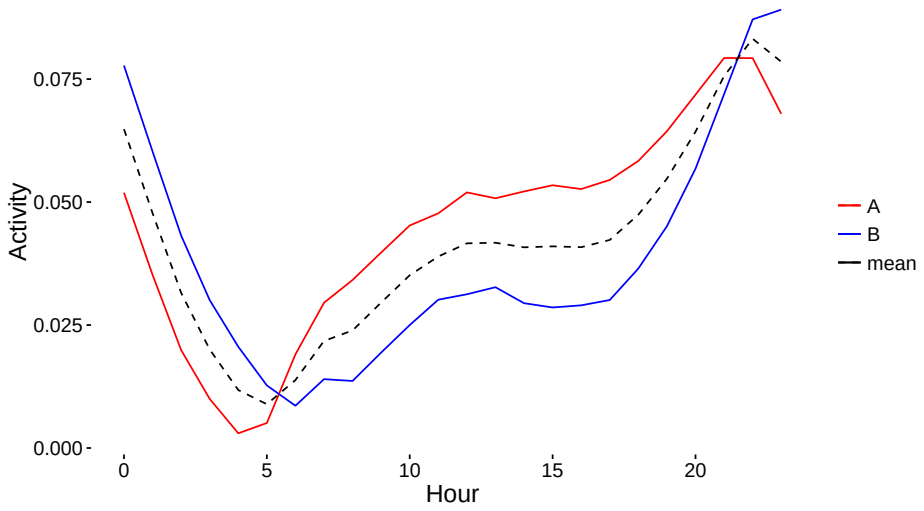


Figure 4.4. Activity patterns corresponding to the two population groups. The red line, **A** corresponds to the daily activity pattern of a population with regular working hours. The blue line, **B** corresponds to the other group who stay up until later in the evening and wake up later as well. The dashed line marks the average activity of all counties.

Figure 4.4 shows the elements of the 24-dimensional **A** and **B** from Eq 4.14-4.15. By interpreting these patterns as the different average tweeting patterns of two population groups, each $\alpha^{(k)}$ is proportional to the share of people in a county in one population group. A possible hypothesis is that the group more active during the daytime corresponds to people who regularly go to work, school etc. on weekdays, thus their daytime is regulated by the earlier wake-up and bedtime indicated in pattern **A**. On the other hand, pattern **B** could correspond to a group where this regulation factor does not exist due to retirement, unemployment or any other reason, which would allow these people to be more active during nighttime and wake up later.

To confirm this hypothesis, I correlate $\alpha^{(k)}$ values with labor force and unemployment estimates from the Local Area Unemployment Statistics (see Methods on p. 62) of the investigated counties. In the 2-dimensional case, these combined values of $\alpha^{(k)}$ do not correlate with employment (0.38 ± 0.03) or unemployment (-0.32 ± 0.02) better than previous activity measures from single dimensions from Figure 4.1. However, by using all dimensions, correlations of 0.46 ± 0.02 and -0.34 ± 0.02 are found for employment (see scatterplot in Figure 4.5) and un-

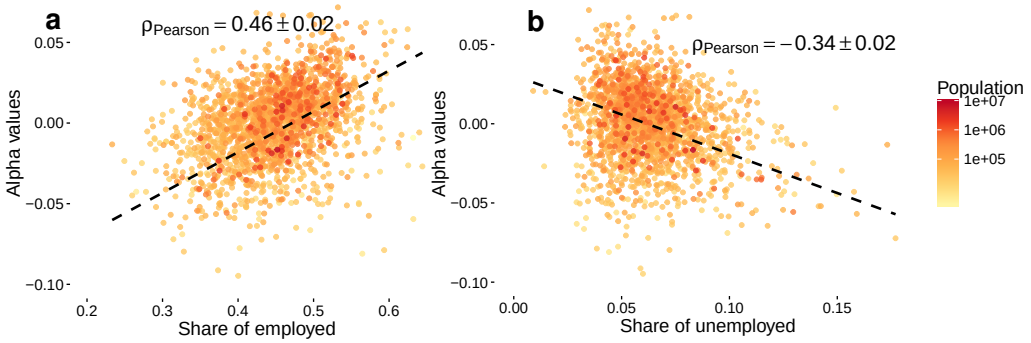


Figure 4.5. Scatterplots of 24-dimensional projected $\alpha^{(k)}$ values with employment and unemployment

employment, respectively. For the employment, this is an improvement to that of the single dimensional correlations, while it is not for the unemployment. A possible interpretation is that a stricter daily rhythm is imposed upon those who are employed, as such, the characteristics of their activity curves mean a stronger overall pattern than that of the unemployed. Nevertheless, the result shows that high a $\alpha^{(k)}$ is significantly bound to higher employment, and lower unemployment rates, and that the overall shape of the activity timeline can give us more information than just using one feature of a whole day. The similarity of the regional distribution of $\alpha^{(k)}$, unemployment and employment rates are visualized on the three maps of Figure 4.6.

These results are in line with previous research carried out for Spain in [209], where share of Twitter activity during a window of the morning hours (8-11am), afternoon hours (3-5pm) and of the night hours (0-3am) correlated significantly with unemployment rates among 25 to 44-year old inhabitants of Spanish administrative areas. High morning and low night activity indicated lower unemployment rates, which is in correspondence with the obtained correlations. Although in Spain high afternoon activity correlated positively with unemployment levels, this phenomenon cannot be observed in the US. Due to the bias in the age of Twitter users towards younger age groups [81], the calculated county activity patterns are not representative of the whole population. I believe that this model could be improved by incorporating labor force data detailed by different age groups.

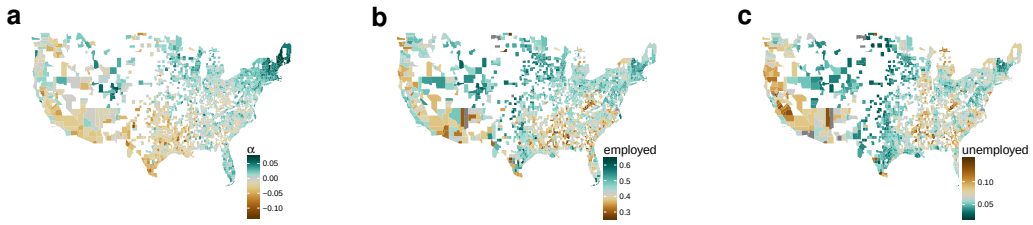


Figure 4.6. Map of $\alpha^{(k)}$, employment and unemployment levels. Regional similarities are visualized by plotting **a** $\alpha^{(k)}$ measures, employment **b** and unemployment **c** on a US county map. Blank counties did not exceed the 1800 tweet threshold described in the Methods section.

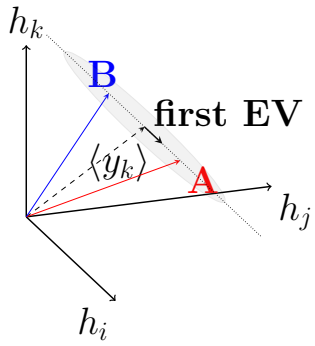


Figure 4.7. Schematic figure of linear model. Linear representation of the datapoints by **A** and **B** corresponding to the two universal patterns later identified as the active and inactive patterns, the dashed line showing the mean value of activities. The bold vector is the direction of the first eigenvector of the covariance matrix **C**. h_i, h_j and h_k represent three arbitrarily chosen axis corresponding to different hours i, j and k of the day.

That correlation with unemployment is significantly lower than correlation with labor force share of the population can be related to the fact that the share of employed should overlap more with the population exhibiting the "working" pattern **A**, whereas officially registered unemployed people are not distinguishable in this context from those who are on a maternal leave or are retired. There might be other inherent reasons, for example, the more flexible working hours in the creative industry that limit the power of such a simple model explaining the employment patterns of a geographical area.

4 Conclusions

In this chapter I analyzed an extensive collection of geolocated tweets originating from the United States between January 2014 and October 2014. Each tweet was assigned to a county, then daily tweeting activity patterns were aggregated for a typical weekday, and I investigated the extent to which hourly activities correlate with employment or unemployment levels. Daily activity patterns were then modelled as being the superposition of two universal patterns, thus aiming for a simple linear approximation of the dataset. By minimizing the squared error of the estimations, I obtained that the difference of the two patterns should be parallel to the first eigenvector of the covariance matrix of the dataset and that the mean of the data should fit on the line when selecting only 2 dimensions, and when using all 24 dimensions of the data as well. The set of eigenvalues of the covariance matrix in both cases confirmed the validity of the linear model, which captured most (0.99, 0.52) of the variance in the dataset. Whereas in the 2-dimensional case the first eigenvector pointed to the direction where 1pm activity was increased, and 12am activity decreased, in the 24-dimensional case, it had positive elements during the daytime hours (6am-8pm), and was negative during the most of the night (9pm-5am).

By projecting county activity patterns onto these lines with the mean as the origin, I obtained a measure for each county that indicated the extent to which the tweeting pattern of a county resembles that of the first eigenvector. This measure has been shown to correlate significantly with county labor force shares and unemployment rates, though in the 2-dimensional case, these correlations could not enhance the performance of the single hourly correlations. Using all 24 dimensions, I obtained a better Pearson correlation of 0.46 ± 0.02 and -0.34 ± 0.02 for employment and unemployment, respectively. The signs of the correlations indicate a relationship where counties exhibiting a higher tweeting activity during the daytime (6am-8pm) have higher employment and lower unemployment rates, and counties with increased night activity can be related to lower employment and higher unemployment rates. These correlations show, that even though Twitter population is biased towards younger age groups, and employment data was considered for all age groups, the

underlying relationship between daily activity patterns and employment data can be captured with plausible outcomes.

The results thus showed, that by analyzing a relatively sparse publicly available geolocated dataset, a very simple model can explain to a significant extent such an important socio-economic indicator as employment/unemployment. The model could even be further improved by incorporating detailed data for different age groups or other datasets from either traditional or digital sources such as mobile traffic data. It would be worth to investigate whether dynamic changes of activity patterns over time can follow employment trends. This kind of analysis would allow policy makers a better insight into the processes connected to employment phenomena, and could form the basis of future datasets, where problems could not only be identified based on officially registered unemployed people, but also on a basis of the digital footprints people leave on different platforms.

5

Urban land use detection

Cities are constantly evolving complex systems. Detection methods of land use distribution have to keep pace with their changing landscapes. Traditional analysis relies on surveys refreshed at most yearly. However, because of the widespread use of mobile devices, cell phone activity measurements can be used as sensors for the functional clustering of urban districts. These activity-based proprietary measurements have lately been complemented by publicly available geosocial data that enables a content-aware analysis. In this work, I analyze the relationship between conversation content and functional spatial clusters of cities with a double dataset approach. I look at the differentiating power of the content of local conversations in activity-driven land use detection that is based on mobile phone records. Three metropolises, London, New York City, and Los Angeles are analyzed separately and in comparison to each other. I show that the share of words with a similar temporal pattern to that of local mobile activities is different across cities, as well as between functional clusters. Moreover, conversational content can differentiate both functional clusters of a single city, and similar locations of the same function across many cities, like business areas that otherwise have a common temporal heartbeat. Words related to activity types are then identified as the most important features emerging from the content-based, data-driven classification.

The material presented in this chapter appeared in [4].

1 Introduction

1.1 Background

As already mentioned in Chapter 1, the global population increase drives urban expansion faster than ever before, as confirmed by the survey of the United Nations Department of Economic and Social Affairs [90]. This accelerated urbanization process means that the key to the solution of many global problems lies in making various aspects of cities more efficient. Such aspects range from planning optimal urban public transportation systems to dividing urban space between several important activities such as residence, commerce, industry or recreation. Thus, understanding and modeling urban mobility and land use patterns are of major importance [193].

Traditional methods of detecting urban land use have several shortcomings. Offical records might be non-existing and they are difficult to keep up-to-date. Moreover, actual land use might differ from the intention of authorities that makes these records unreliable. Checking the real state on site often includes conducting costly individual surveys, that suffer from the usual drawbacks of being slow or having low or inconsistent response rates. With the recent availability of satellite imaging, land use can be automatically inferred using digital image processing techniques in combination with Geographic Information Systems (GIS). However, the frequency of GIS imaging is not able to uncover any dynamic aspects of land use. As such, land use records are hard to keep up-to-date with the rehabilitation projects and the spontaneous reorganizations of the city life.

Apart from the time delay, the lack of actual social aspects of land use is also a disadvantage of such static methods. Social aspects include how exactly people interact with these urban environments, which is a crucial piece of information helping the work of analysts and decision-makers alike. Recently, the availability of large-scale mobile phone datasets enabled the study of human mobility at a previously unprecedented scale. Mobile phone data can be regarded as a certain type of collection of digital traces that people leave behind with their devices. These traces then act as a human sensor network [39, 212], that allows an almost

real-time analysis as opposed to the processing of census or GIS data, or the use of large, paper-based surveys. Therefore, *quantitative or activity-based analysis* of mobile communication records is the core of many urban studies.

One branch of these studies focusing on inferring land use from mobile phone traces builds on supervised learning methods. This means that they try to detect land use patterns using labeled ground-truth data. These are in some cases official records, in other cases crowdsourced points of interest. Dashdorj et al. [213] used both types of data in their paper, and they found that official land use records are harder to predict than crowdsourced OpenStreetMap data from mobile timelines with the applied machine learning methods. This suggests that OpenStreetMap is more capable of following the updates and reorganizations of city life. A Random Forest Classification trained on zone labeling yields reasonable clusters in the article of Toole et al. [191]. Another paper from Pei et al [192] detects land use clusters by semi-supervised clustering, using expert labeled data to train the algorithm.

Supervised methods still require predefined labels, that alone might introduce errors into the detection process [191, 183]. Therefore, unsupervised learning methods are also widespread in this area. A milestone of this approach was using the eigenvector decomposition of the spatiotemporal mobile phone activities [188]. The first few eigenvectors capture the most important patterns in Rome, showing typical signatures of daily city life. Since then, k-means clustering [190, 194, 186], hierarchical clustering [197], Independent Component Analysis [214] or Latent Dirichlet Allocation [215] have also been applied to decompose urban spatiotemporal dynamics into the most typical signatures. The automatically provided land use clusters uncovered by these unsupervised methods align well with ground-truth data. Thus, they could be used for almost real-time monitoring of actual patterns and trends, such as crime prediction [202], ambient population density [216, 208], or urban greenspace usage [218].

However, mobile phone datasets are often proprietary and hard to obtain. Therefore, there has been significant interest in channeling freely available data sources into the above methods. Such data sources are the different location-based social networks such as Twitter or Foursquare. Many studies exploit the freely available Twitter stream introduced in Chapter 1, for example, spectral clustering of Self-

Organized Maps of urban Twitter messages provides similar land use information to that of mobile phone clusterings [219, 211]. Latent Dirichlet Allocation also proved to be useful on Twitter data by Ríos and Muñoz [215]. Steiger et al. [221] argue that Self-Organizing Maps that construct the division of the urban space based on the data itself solve the problem of modifiable basic units present in studies using arbitrary divisions or administrative units. Social networks and people's travel behavior are connected to each other [222], as well as social network activity in certain ranges of a day and land use types [223]. Apart from the spatiotemporal aspects, social media data can be reflective of the content of the local communication, that deepens the understanding of the land use signatures [221, 224]. For further reference on using mobile phone datasets in urban sensing, see the extensive reviews of Jiang [225] and Blondel [226].

1.2 Focus of the study

Despite the coexistence of various methods and datasets, location-based content-aware studies are not only sparse, but the cross-validation of the findings with census data or with mobile phone records is also underrepresented. The study of Lenormand et al. [227] suggests that the signatures found in mobile and social media datasets correlate significantly. Because only a few studies addressed the cross-checking of different data sources, those possessing the proprietary mobile phone data and obtaining the free social media data are able to provide meaningful comparisons. Therefore, researchers for whom only social media data is available are able to assess the validity of their results. To bridge this dual landscape of urban studies, this chapter analyzes the connection between conversation content from social media data and functional spatial clusters of cities based on mobile activity timelines.

This work is based on a double set of geo-traces of communication measurements: first I use the results of a previous quantitative analysis of mobile phone records of Grauwin et al. [194], and second, I add qualitative aspect by analyzing public Twitter messages of the same locations. Geo-tagged messages of three different cities are aggregated into the same spatial grid as used in the former work of land use detection. These three cities include London, New York City, and Los Angeles,

because their mobile call data and land use clustering is available, and they share English as a common language. While datasets of mobile phone records are detailed in [194], geosocial measurements are discussed in Materials and methods. Fig. 5.1 illustrates the explanatory power of the geosocial content with three examples.

In what follows, first, the general relationhsip between original functional clusters and the content of communication is investigated. Then, the correlation between the original activity profiles and the word timelines corresponding to the same areas is examined. Next, I look at the intra-urban distinctive power of words by trying to recover the original functional clusters strictly from local word frequency information. Finally, I analyze the ability of communication content to distinguish inter-urban classification of pixels mixed from various cities but from the same functional activity clusters.

Details of the data collection and the used analytic methods are presented in the Materials and methods section. The results are briefly summarized in the Results section and then further discussed in the Discussion section.

2 Materials and methods

2.1 Data collection

For the present chapter, the freely available 1% sample of the Twitter API described in Chapter 1 is used. With the help of the sample filtering parameters, geotweets were queried between 1 January 2015 and 31 December 2015 falling within the three bounding boxes of three distinctive metropolises: London, New York City, and Los Angeles. Table 5.1 lists the corner coordinates of each bounding box and the number of geotweets originating from these bounding boxes for the three investigated cities. These cities have been selected for complete comparative analysis based on multiple features. First, we can compare high volume Twitter measurements at the same locations to the results of [194] from these same cities. Second, they all share English as a major tweeting language, which allows direct comparison between them.

city	SW long	SW lat	NE long	NE lat	number of geotweets
London	-0.510	51.286	0.334	51.692	6,959,417
New York City	-75.997	39.505	-71.856	42.180	10,769,664
Los Angeles	-124.409	32.534	-114.130	42.009	6,520,472

Table 5.1. Data collection bounding boxes given to Twitter queries. The table contains the data collection bounding box coordinates given to Twitter queries (SW - southwest, NE - northeast corner) and number of collected geotweets for the three cities analyzed.

2.2 Data aggregation, notations

Because Twitter messages are too short for individual assessment with traditional topic modeling analysis, the data is aggregated both in space and in time domains. Spatial aggregation uses the same pixel resolution on square grids as in [194] that

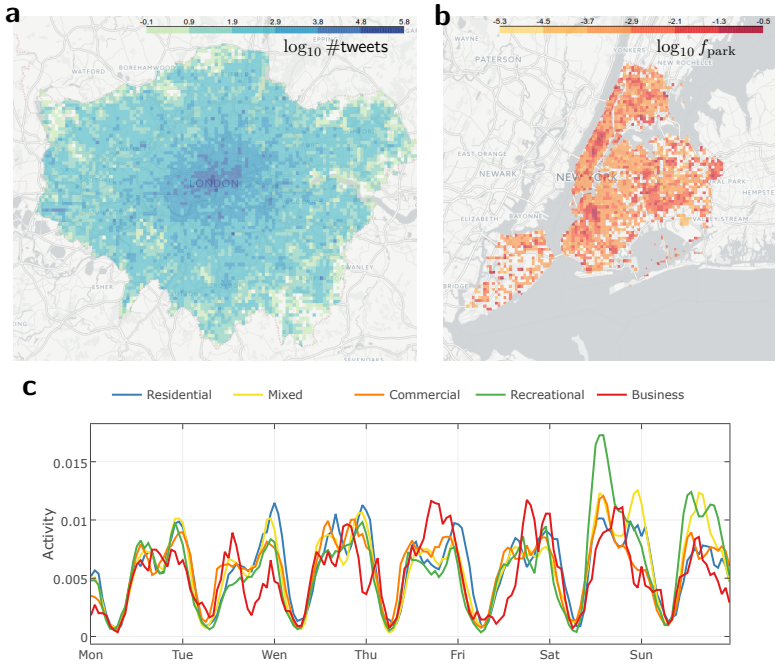


Figure 5.1. Examples showing how urban communication activities represent geo-social aspects. The subfigures show examples of how Twitter data segments the spatial and time domains in three cities. Spatial grids are the same as used for the study in [194]. Uncolored pixels contain no data. (a) Spatial distribution of the number of Twitter messages in London aggregated over the entire collection period. (b) Log frequency heatmap of the word 'park' in New York City. (c) Activity timeline of the word 'run' across different land use clusters in Los Angeles aggregated into the hours of a week.

enables assigning a functional label to each pixel, that I obtained from [194]. That is, for each city, a 0.5km×0.5km grid has been used, and the (lon,lat) coordinate pairs of the tweets were sorted into these spatial grids. Each tweet is assigned to a time bin, according to the hour of the week the tweet had been sent from (beginning with 0 and ending with 167). Because small spatial pixels do not contain enough words to retrieve meaningful weekly word timelines, the average timelines of the words are calculated for entire functional clusters. An example of the cluster timelines for the word 'run' can be seen in Figure 5.1c.

To calculate word frequency distributions in the pixels, I parsed the text of the tweets into words. First, URLs, usernames beginning with @, and hashtags beginning with # were removed. Then texts were broken into words using the Twitter-specific tokenizer of [228] and words were lemmatized with [229, 221].

For the word frequency analysis, I created word (lemma) frequency matrices W^c for each city c, where the elements $W^c_{ip} = n^c_{ip}$ represent the number of occurrences of the ith word in the pth pixel of city c. Before any further processing, pixel-wise word counts were saved for normalization purposes, denote these by n_i^{norm}. Stopwords were then removed with the help of [231], and words that were not frequent enough ($\sum_p W^c_{ip} <$ min wf^c), or did not occur in enough pixels of a city ($\sum_p \Theta(W^c_{ip}) <$ min wv^c) were cut out. I also omitted pixels that did not contain enough words ($\sum_i W^c_{ip} <$ min pf^c) or enough types of words ($\sum_i \Theta(W^c_{ip}) <$ min pv^c), where $\Theta(x)$ is the Heaviside function. Table 5.2 shows the filtering thresholds and the number of remaining pixels and words.

city	min wf	min vw	min pf	min pv	num words	num pixels
London	100,000	500	30,000	1,500	2,979	979
New York City	100,000	500	30,000	1,500	3,097	969
Los Angeles	20,000	500	20,000	700	2,578	2,452

Table 5.2. Filtering conditions for the word-document matrices, and number of remaining words and pixels after the filtering process. The table contains the filtering conditions for the word-document matrices, and number of remaining words and pixels after the filtering process. Columns: *min wf* - minimum overall word count per pixel, *min wv* - minimum types of unique words in a pixel, *min pf* - minimum overall count of a word in all pixels, *min pf* - minimum number of pixels a word appeared in, *num words* - number of remaining words, *num pixels* - number of remaining pixels.

Thus, after filtering and normalization, $\mathbf{W}^c$ matrices look like the following (the c index of the elements is left out for better readability):

$$\mathbf{W}^c = \begin{pmatrix} & \vdots & \\ \dots & n_{ip}/n_i^{norm} & \dots \\ & \vdots & \end{pmatrix}.$$

From the clustering of the mobile networks, the partitioning of the pixels p is already given. Let's call these partitions or clusters K_k^c, where c denotes the city, and where $k = 1, \dots, 5$ in the case of New York City and Los Angeles, and $k = 1, \dots, 6$ in the case of London. Each cluster is a set of pixel indices according to the original map. Let's consider now only the filtered pixels as the elements of these sets.

Define the following measures connected to the ith word:

- city-wide average
$$E_c(i) = \frac{1}{|P|} \sum_{p=1,\dots,P} n_{ip}^c / n_i^{c,norm},$$
- city-wide standard deviation
$$\sigma_c(i) = \sqrt{\frac{1}{|P|} \sum_{p=1,\dots,P} \left(n_{ip}^c / n_i^{c,norm} - E_c(i) \right)^2},$$
- cluster-wide average
$$E_{K_k^c}^c(i) = \frac{1}{|K_k|} \sum_{p \in K_k^c} n_{ip}^c / n_i^{c,norm},$$
- z-score in a cluster K_k^c
$$z_{i,K_k^c} = \frac{n_{ip}^c / n_i^{c,norm} - E_c(i)}{\sigma_c(i)}.$$

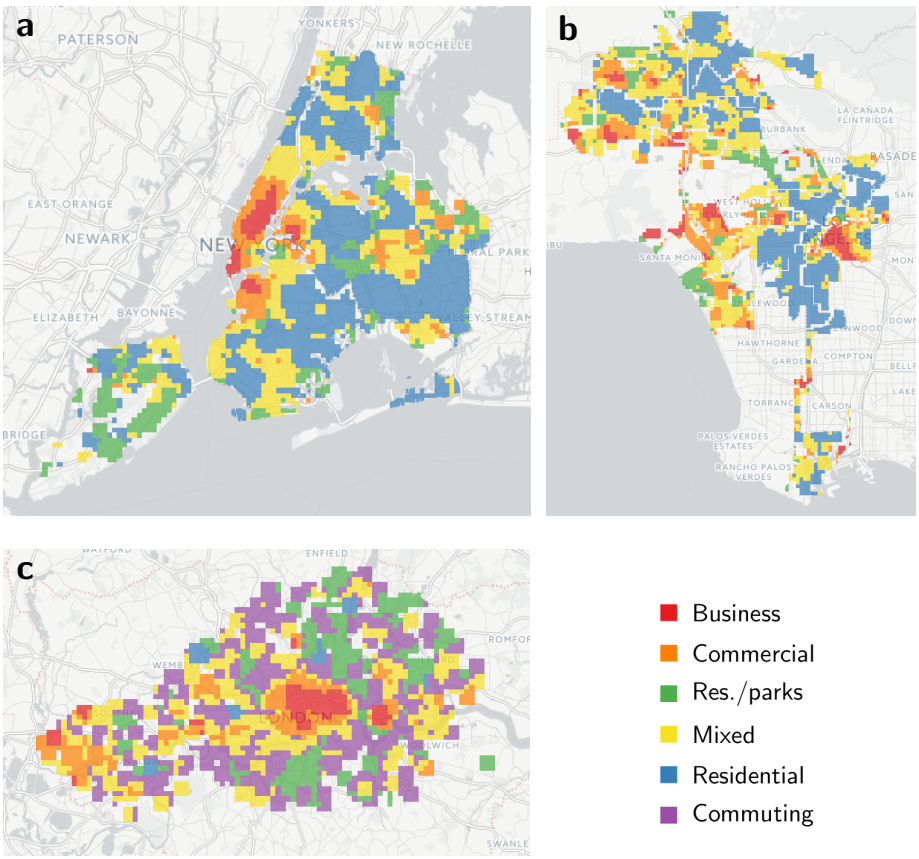


Figure 5.2. Land use clusters on the spatial grid based on the clustering of mobile activity timelines. The three figures show the land use clusters on the spatial grid based on the clustering of mobile activity timelines for the three cities: (a) New York City (b) Los Angeles (c) Greater London [194].

Machine learning

For every pixel p, the column p of the $\mathbf{W}^c$ matrix that contains the word frequencies defined above is taken. Then a Random Forest classifier [232] is trained for each cluster in a city with 10-fold cross-validation, such that the true labels as if all pixels that are in clusters had label 1, and all of the others label 0. The number of estimators in each classifier was 10,000, the maximum number of features used 500, and the maximum classifier depth 4. The goodness of the clusterings is assessed with the Area Under Curve (AUC) indicator [233]. The AUC score gives the probability that the corresponding classifier ranks a randomly chosen pixel with label 1 higher than a randomly chosen pixel with label 0 [234]. AUC is equal to 1

if the classification is perfect, that is, the classifier perfectly recovers the original labeling in the test set. However, it is equal to 0.5, if the classifier behaves like a totally random classifier assigning 1 or 0 with a probability 0.5 to the pixels. The closer to 1 the AUC score is, the better the classifier performs.

2.3 Top feature selection

First, the most significant words of a cluster k in a city c were obtained by taking the words with the highest z-scores.The set of all words that are present in the two cities c_1 and c_2 is $\{j|j \in K_k^{c_1} \text{ and } j \in K_k^{c_2}\}$, and it becomes possible to compare clusters k across city pairs (c_1, c_2) with the help of the following two sets:

$$\begin{aligned} T_{c_1} &= \{j|z_{j,K_k^{c_1}} > 0.5\}, \\ T_{c_2} &= \{j|z_{j,K_k^{c_2}} > 0.5\}. \end{aligned}$$

The Jaccard-index between these sets is given by:

$$J_k^{c_1,c_2} = \frac{|T_{c_1} \cap T_{c_2}|}{|T_{c_1} \cup T_{c_2}|}.$$

3 Results

Quantitative activity timelines: mobile phone vs. word usage patterns

To measure the similarity between these two weekly timeline shapes, I correlated the quantitative activity patterns of words to those of the mobile-activity-based functional clusters. Figure 5.3a shows the fraction of words that are not correlated, and Figure 5.3b shows the fraction of words that are highly correlated to the timeline shapes. In New York City, the non-correlated word share is consistently lower throughout all clusters than in the other two cities. Moreover, the share of highly correlated timelines is at the same time much bigger throughout all clusters.

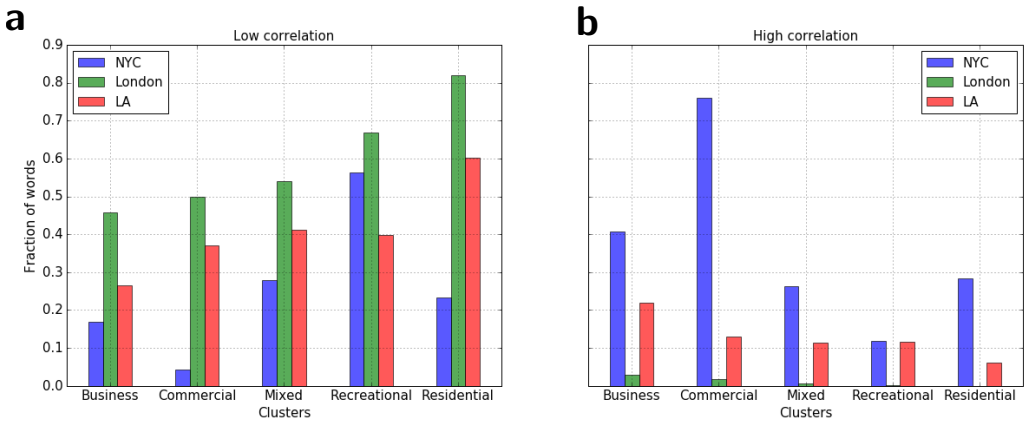


Figure 5.3. Fraction of words inside functional clusters based on timeline correlations with mobile activity. The figures show the fraction of words whose timelines are or are not in alignment with the mobile timeline of the functional cluster. (a) Fraction of words whose timeline correlates with $|C| < 0.2$ with that of the mobile-activity-based cluster. (b) Fraction of words whose timeline correlates with $|C| > 0.5$ with that of the mobile-activity-based cluster.

This effect is not as pronounced, but still existing in the case of Los Angeles. Note how business and commercial clusters show a higher alignment with their word timelines when considering both Figure 5.3a and Figure 5.3b in the case of New York City and Los Angeles. However, word timelines in London align much less with the mobile activity profiles of their clusters. This is reflected in the high share of non-correlated, and in the extremely low share of highly correlated word timelines.

3.1 Qualitative analysis: intra-city distinctive power of words

It is also interesting to investigate word distributions in different clusters of the same city and their cluster-level differentiating power. Here the focus is on the overall frequency of the words spatially aggregated into pixels - like above - for the whole data collection period (see Data aggregation, notations section in Materials and methods for details). Based on the pixel word frequencies, a Random Forest classifier [232] is built for each city (see the Machine learning section in Materials and methods) to learn the cluster labels.

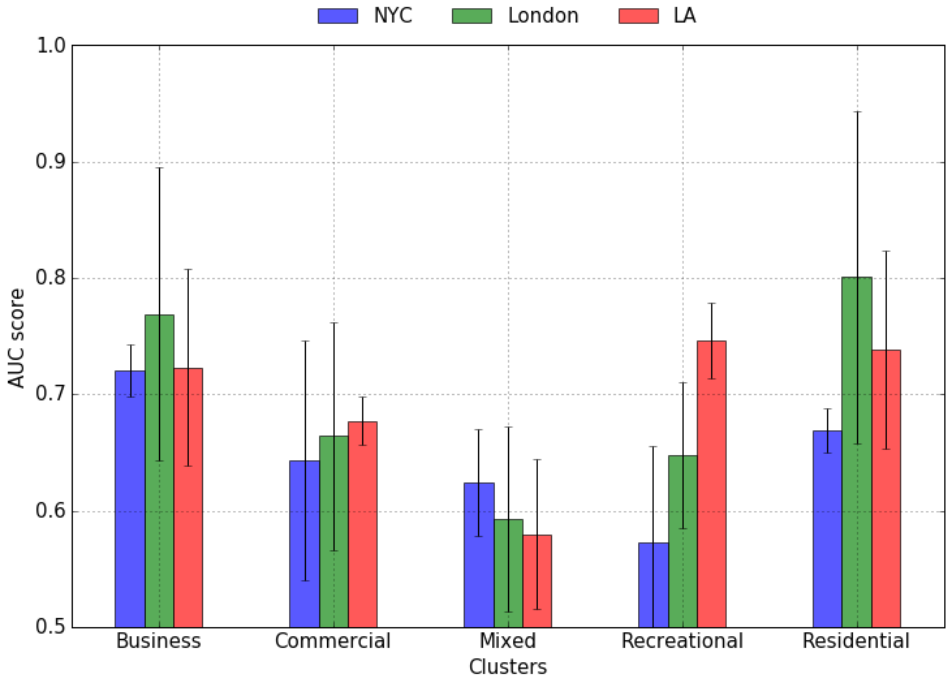


Figure 5.4. Mean AUC scores of the Random Forest Classification. The figure shows the mean AUC scores of the Random Forest Classification that tried to separate different cluster labels based on the Twitter language. Results are based on overall word frequencies of city pixels representative of the predictability of the original mobile phone activity clusters in each of the three cities. Error bars are calculated using the standard deviation of AUC scores for the clustering of different data types.

Figure 5.4 shows the performance of the classifiers broken down to city and cluster labels characterized by their AUC scores [233]. Altogether, each classifier performs better than a truly random one, which would yield an AUC of 0.5. In general, pixels belonging to Business and Commercial clusters are classified consistently well in all of the cities.

Cl.	LA	London	NYC
1	el cent downtown art start disney ucla district dance create grand drive bay station gallery free shoot lot hall red op garden pumpkin top sport light lead lane space build shot pow	night city street st station big tower market amazing museum square ben working cross palette place coffee british lunch westminster waterloo sky king's lock food lane free pancras theatre visit bank hospital spitalfields office river stick event planet loved building row official meeting international martin job exhibition flower	park central bridge day happy special art video half sale kid enjoy air sunday top shack god train madison pizza hour head hair alert visit supervisor harbor link
2	best drink week west amaze class interest museum heal big hotel blvd enjoy college find nice pretty university glass gold brew auto support visit picture mile room company pm sushi finish talent	i'm greater hpa heathrow uk posted bridge buckingham open airport stratford hammersmith statue regent's herbs we're modern shard art bus cute experience design studio winter walk add finally james style business minutes latest bbc opening waiting tea hello fair corner middlesex	dude square best sign hard event store man fun weekend well williamsburg women baby care staten call red
3	beach i'm monica time venice studio city del click long tail burbank morning universal friend apply recommend view beautiful glendale sign track restaurant pi graffiti sale pacific bowl walk dinner ocean cafe rose dodge beauty stadium full sunset grill star hot wind perfect water dog feature associate serve queen breakfast	mclaren entire birthday common you're lol lewis hey cut academy hell field books greece forest lovebox fields bird sister mother	incident newark catch island nj exit update sb cafe starbucks eat garden liberty matte bike house dinner mark bob jersey blue 100 guy join sister warehouse pair counsel ready shoppe media fly dream yoga office kill gallery high stadium throwback
4	lol fuck shit feel de lmao call ass bitch south talk sleep nigga en ain't damn ur cute stay yeah snapchat tho crazy heart tweet change phone smh money	hill cake today festival work tengo man black meet win wit rugby cast season manor tweet uber premier gorgeous skills cup let's spurs active tfl stunning shoot	bronx queen station work lay year we're latest hall girl river click food fit apply health recommend road sleep current

Table 5.3. Most significant words in the functional clusters of the three cities. Most significant words in the functional clusters of the three cities. See Materials and methods for detailed descriptions on the selection. The four cluster types are: 1: Business, 2: Commercial, 3: Recreational, 4: Residential.

3.2 Qualitative analysis: inter-city comparison of functional clusters

When comparing multiple cities, mobile phone timeline patterns show common weekly rhythms in the pixels forming the Business cluster. Here I demonstrate how content-aware qualitative study reveals well identifiable marks of the different cities' business quarters, as well as any other part of their urban landscapes.

First, multiple Random Forest classifiers are trained to separate pixels belonging to the same functional cluster coming from different cities. The outcome of multiple runs on this classification problem gives an almost perfect AUC (0.99, 1.00) for every land use type. Thus, even the Business clusters of different cities are well separated from each other, when viewed not only from the time domain but from the content-based approach. The most distinctive features of the classifier for the Business cluster are listed in Table 5.4 in decreasing order of importance. The names and abbreviations of the cities (Table 5.4 in bold) are among the top ten most distinctive features. The following features also contain certain city parts or city-specific words ('Brooklyn', 'Hollywood', 'California', 'NJ', 'avenue', 'beach'). Other top features include words connected to touristic activities ('photo', 'posted') or advertisements ('opening', 'click').

ny i'm **york london los** day great latest **angeles** work park time love today we're street it's opening night photo click good nj city posted brooklyn happy **la** california **nyc** fit greater station incident apply recommend tonight avenue de will hollywood center square best morning view united team beach you're

Table 5.4. List of the most distinctive words used for the classification of the Business pixels into different cities. The table lists the most distinctive words used for the classification of the Business pixels into different cities. City names and abbreviations are typeset in bold.

4 Discussion

4.1 Quantitative activity timelines: mobile phone vs. word usage patterns

Since mobile phone datasets are often proprietary and the content of the data is hidden due to privacy reasons, it is important to understand how open-source datasets, such as social media content, relate to mobile activity measurements. Giving activity profiles or mobility trajectories semantic content enriches our understanding of the time-domain patterns that can be observed [224, 235]. The analysis shows that the share of words whose cluster-wise timelines align well or badly with the mobile cluster timelines varies significantly from city to city. It is interesting to see how the two American cities - New York City and Los Angeles - show a greater agreement of the word and mobile timelines than London. Probably, the structure of the two investigated American cities are inherently different from that of London in the sense that the functions present in them are more distinct. The greater overlap between the functions in the London clusters causes a more diverse word pattern that leads to a decrease in the 'harmonization' of the word timelines and mobile phone timelines.

Note that the temporal resolution of the data allows for capturing events only on longer timescales. Twitter, as a conversational platform, is often used to facilitate reporting or discussions of global topics. Hence, tweets that are related to big events such as the London Marathon quickly spread over the entire area of a city. Smaller events that are only of local interest become undetectable due to the sparsity of the data. I used an aggregation over a year's worth of messages to obtain a spatial resolution in the text comparable to that of the mobile phone records. Thus the resulting timelines represent an average or typical behaviour.

4.2 Qualitative analysis: intra-city distinctive power of words

Social media activity [215, 223, 235, 227] or content [221, 224] can segment a city into different land use types. The question is, whether a clustering obtained from a given type of dataset - such as mobile activity - can be validated by another dataset. The extent to which the mobile clustering can be retrieved is characterized by the AUC score of a Random Forest Classifier. In the present chapter, the trained classifier always performs better than a random guess. Even the error bars for the AUC scores, originating from the classification based on clusters of different data types, stay above the random threshold of 0.5. It can be observed that Business and Residential areas perform consistently better at the recognition task (with AUCs higher than 0.7 in 5 out of 6 cases, see Figure 5.4) than the other land use types. A possible explanation can be that the core of a city formed by the pixels labelled as Business has a quite characteristic word pattern because of the city center specific features (e.g. local touristic attractions) that are present there. Likewise, the lack of these specific features can easily identify a pixel as a Residential one. It can be expected that the Mixed land use type is not easily recognizable (having the lowest AUC), since the various topics present in other land use types does not help the classifier to sort Mixed type pixels correctly. It must also be highlighted the Recreational cluster of Los Angeles that is well classified with a small error (AUC above 0.7). Presumably, what sets it apart from the recreational clusters of other cities is that most of its pixels form a relatively compact geographical area along the city's famous beach. This separates it well both in the time domain and the topic space.

By looking at the words that highlight a certain cluster of a city in Table 5.3, it is possible to assess the content that sets the functional clusters of cities apart can be assessed. It is expected and confirmed from the word analysis, that geographically bound landmarks of cities appear in the clusters in which they are physically present. For example, note the words 'UCLA', 'Disney' in the Los Angeles, 'Big Ben', 'King's Cross', 'Tower Bridge', 'Westminster', 'Waterloo', 'St. Pancras' in the London, 'Central Park', 'bridge' and 'harbor' in New York City Business clusters,. The Commercial cluster in London ('HPA', 'Heathrow', 'Stratford', 'Shard' etc.) and the Recreational cluster in LA ('Long Beach', 'Burbank', 'Glendale') also

show the same landmark-specific top words. This phenomenon is most prominent in the Business clusters, which can relate to the fact that Business cluster pixels are mostly located in the core of the cities where the density of landmarks and touristic attractions is higher. However, not only landmarks, but activity-specific words are also in alignment with land uses. For example, the words 'hotel', 'university', 'event', 'store' in the Commercial, and 'view', 'beautiful', 'stadium', 'grill', 'water', 'forest', 'bird', 'garden', 'yoga', 'shoppe' in the Recreational clusters. Again, these words belong to city-specific activities, like 'water' being representative of Los Angeles, 'forest' of London, while 'yoga' and 'shoppe' of New York City. The Los Angeles Residential cluster is marked by a higher presence of swearwords. These activity-specific words have a general meaning, thus, they can enhance the performance of land use detection algorithms relying solely on data or call volumes. It is also worth noting that content may enable a better resolution in the determination of functional spatial clusters. For example, poorer or richer residential neighborhoods might look similar according to normalized averaged mobile activity, but conversational content could reveal the differences between them.

4.3 Qualitative analysis: inter-city comparison of functional clusters

Using the content-aware dataset the following questions can be asked. First, is it true that the Business clusters of these three cities cannot be distinguished from each other? Second, do business clusters share a common temporal heartbeat as well as a common vocabulary? One of the most interesting results of the previous study [194] was the high level of similarity between Business clusters' weekly activity patterns in three fundamentally different cities. Actually, after the joint clustering of all of their pixel timelines, business areas emerged as the single dominant cluster, and the clustering did not show any other commonly recognizable functional pattern. To answer the above questions, I performed a new training on the word frequencies (disregarding the time domain again) using the pixels labelled as Business from all three cities. The results show a clear difference between the cities, as almost all (AUC $\approx$ 1) Business cluster pixels are correctly categorized

into their own cities. This suggests that the content appearing in the Business functional clusters of different cities is remarkably different.

The features that separate the three Business clusters are listed in Table 5.4. The name of the cities and their abbreviations dominate the first ten places of this list, which suggests that the content separates mostly along the geographical names. But some features connected to tourism and certain landmarks ('park', 'station', 'street', 'beach') are also present, which indicates that apart from geographical names, certain functions of Business pixels also set the cities apart. Thus, even though the pace of work - measured through activity timelines - resemble each other across different cities, the content is truly unique in a certain Business cluster. Further analysis revealed that cities are recognizable with a similar level of certainty not only based on their Business zone traits, but also based on the relevant content of conversations in other types of detected land use clusters.

Is must be noted here that the generalization of results from the comparative analysis of geolocated Twitter user behavior to the general population is not a trivial task. Twitter users are not representative of the population neither in the US [79] nor in London or in the UK [47, 84]. Mobile phone records can also contain sampling biases across different age groups, because younger people more often have Internet subscriptions with larger monthly cap, and the population is not uniformly distributed or uniformly active within different areas of the cities. These biases can also be present in the aggregated geographical analysis, for instance, because of segregated housing issues. Inferring demographical attributes of Twitter users and especially mobile phone users is in itself a highly demanding task [50, 85, 237]. Nevertheless, Twitter remains the most important source of publicly available geolocated textual data, and this is the most state-of-the-art method by which conversational content can be accessed in different areas of a city.

5 Conclusion

In conclusion, this chapter focuses on validating and enriching mobile activity based land use clustering with the help of publicly available Twitter messages. Through the evaluation of the complementary datasets, it is also possible to gain some

insights into the similarities and the most differentiating features of human activity traces in three metropolises. After the spatial and temporal aggregation of tweets, I first analyzed how word frequency timelines align with mobile activity timelines of functional clusters. I found that the share of words whose frequency moves similarly to that of mobile activities is not the same across the clusters of different cities and that it also differs from cluster to cluster. With the help of Random Forest classifiers, I determined how much of the functional clustering can be retrieved in a certain city based only on words. The performance of the classifiers was the highest for the Business and Residential clusters and was significant in all cases. Words responsible for the separation of city clusters were typically found to be geographically bound landmarks and city-specific expressions, but topics related to the detected cluster functions also appeared as differentiators. Pixels belonging to the same functional cluster have also been separated into different cities, which could be done with almost 100% certainty. Finally, I analyzed the most distinctive words that enabled the separation of cities among pixels belonging to the same functional cluster.

Here, the focus was only on the local typical word uses, because aggregation and filtering hide any anomalies that can happen due to special events or emergencies. The approach could be used for event detection: mining for patterns that are different from typical at a given location or at a given time. Using words that are distinctive of a certain type of cluster could enhance the performance of existing land-use detection algorithms. The added dimension of conversational content could also lead to a greater resolution in functional clustering by breaking down areas similar in call or data volume profiles but differing in local focus topics. I believe that this work can lead to further investigation of connections between various urban descriptors derived from independent data sources.

6

Conclusion

In this thesis, I aimed to understand and model complex human behavior. This understanding is based on the recent availability of rich data sources such as mobile call records and social media data. Here, I used mobile call records, billions of the messages of the online social network Twitter, and a detailed historical database of the voting data of United States counties for explaining aggregate patterns. In Chapter 1, I gave an extensive review of the background of using geographically tagged social media data for predicting real-world outcomes, and I also introduce Twitter and the structure of its messages in detail. I also discussed the potential biases and drawbacks using the freely available Twitter data. Despite these issues, I still believe that Twitter is able to provide meaningful insight into various aggregate phenomena, as I show in the chapters following Background. The end of Chapter 1 introduces the reader into the literature on urban scaling, that is the underlying theory in Chapters 2-3 of this thesis. In the following paragraphs, I would like to give a brief summary of each of the chapters of this work.

First, in Chapter 2, I use the urban scaling framework to explain the historical voting patterns of metropolitan areas in the United States. First, I show how votes cast for the Republican and Democratic parties in the 2016 presidential elections of the United States fit the laws of urban scaling. This fit reflects previous observations about bigger cities voting more for Democrats. In the urban scaling framework, this is equivalent to the fact that the votes given the Democratic party show a

superlinear scaling as a function of the voter turnout in the metropolitan areas. Moreover, I showed that the scaling holds for historical election results beginning from 1960, with the scaling exponents of the two parties being dependent on each other. The dependence follows from substituting the scaling relationships into the conservation of the voting probability when summing up for all parties. Thus, the measured exponents fit well the theoretical expectation from the scaling laws in each election year, and out of the two exponents, one is enough to explain the other. Then I show that similarly to an urban scaling model of Gomez-Lievano et al., the intercepts and scaling exponents are also connected by a linear relationship using all fits from the historical election results. Therefore, to explain one election, it is enough to model the scaling exponent of one of the parties, and the scaling exponent of the other party and the intercepts of the fits can be calculated from that single exponent using the universal parameters characterizing the historical election process. I then tested the distribution of the normalized logarithmic deviations from the party scaling curves, and I found that only the deviations for the Democratic party follow a lognormal distribution, that corresponds to urban scaling phenomena. Therefore, urban scaling for the Republican party is just the result of the probability conservation process. Finally, I was able to explain the voting behavior for the superlinearly scaling Democratic party by applying the economic model of Gomez-Lievano et al. to voting processes, where a number of values or issues have to be tolerated by an individual voter in order to vote for the party. Bigger cities make their inhabitants more tolerant towards certain minorities or beliefs though processes of social contagion and growing cultural diversity. Using this fact, I successfully explained the growing gap between the scaling exponents of the two main US parties by showing that the number of issues or values a voter has to accept has increased over the years. The model seems to be applicable to other voting processes such as the 2016 EU referendum in the United Kingdom, where the votes cast for the Remain opinion also exhibit a similar superlinear phenomenon as the Democratic voters. Here, the model underlying the results can explain the famous "immigration paradox", which means that communities having more diverse ethnic backgrounds voted more for Remain, that was the more immigration-friendly option.

Second, in Chapter 3, I apply the urban scaling theory to the language of geographically tagged social media messages. In this chapter, I showed how single

word frequencies in the different metropolitan areas of the United States obey the urban scaling laws. While the total word volume and the total tweet volume is also scaling slightly superlinearly, the exponents characterizing their scaling laws are less than exponents for human interactions from previous literature. Those words that follow the same exponent than that of the total word count, are mostly coming from a core vocabulary of the language, such as the stopwords "the" or "he". Also, words with the most super- or sublinear exponents have a meaning that is representative of the super- or sublinear phenomena from other urban scaling measures. Therefore, the extent of agglomeration in geographical space has a quantitatively well measurable effect on word choice in the language of social media. At the end of the chapter, I also show that the number of distinct words scales sublinearly with the population, which corresponds to Heaps law. This demonstrates how there is a decreasing marginal need for new words as there is a growing number of people, thus, a growing number of tweets in a certain area.

Third, in Chapter 4, I develop a new numerical measure that correlates well with employment levels of geographical areas. This measure is created from the daily activity profile of geographical areas through a linear model. The main idea is that the aggregated daily activity timelines are a linear combination of the average activity timelines of two groups of people. One of these groups has a regular daily temporal structure, imposed upon them by either work or school, and the other group lacks this structure because of being unemployed or otherwise inactive. I used geolocated Twitter messages aggregated into daily patterns within the counties of the United States. Then, I decomposed the detected patterns into the linear combination of the daily patterns of the two groups by minimizing the squared error between the data and the model. As a result, I obtained two base patterns for the active and inactive group of people. The shape of these patterns reflects being active or inactive, with the active pattern corresponding to people getting up earlier and going to bed earlier, and the inactive pattern exhibiting a shifted timeline with getting up later and ceasing activity also later in the night. The coefficient for each county, that gives the mixing ratio of the two base patterns, correlates significantly with county employment level. Thus, instead of measuring the relative activity of a chosen time window from a day, I gave a measure capturing the whole shape of the daily pattern observed in a geographical area correlating with employment levels.

Finally, in Chapter 5, I investigate how urban land use is connected to the context of social media messages sent from areas with differing land use. The segmentation of urban space based on observed call patterns or social media data with geolocation information already has an extensive literature. But this literature is only based either on call activity data that does not focus on the context of the land use patterns, or social media data alone, that lacks the granularity of the call datasets. In this chapter, I presented a comparative study of three cities, London, New York City, and Los Angeles using call data based land use clusters and geolocated messages from the social media platform Twitter. I found that word timelines aligning well with the activity timelines, or words that are relatively more abundant in an area reflect the social context of land use within most clusters. Moreover, I built a machine learning model trying to capture differences in the word usage within different clusters of the same city, and within the same cluster of different cities. I found that social context is capable of reproducing the land use clusters to a significant extent based on a goodness measure of the machine learning classifiers. I also found that while activity patterns alone cannot tell the difference between the Business clusters of different cities, the context-based machine learning is able to set these clusters apart. This approach is useful for determining the extent to which patterns detected in proprietary call data are recoverable from freely available social media data sources.

In conclusion, I attempted to show multiple sides of the applicability of large-scale data sources for understanding aggregate human behavior. I used both the spatial, temporal and textual features of my datasets, and I extended and developed new models for the explanation of collective phenomena measured in the data. I hope that this work provides valuable insights and a basis for future research in the interdisciplinary community that tries to understand complex social phenomena.

Summary

In this thesis, I aimed to understand and model human behavior with tools from complexity science. There are already several studies that use models borrowed from physics such as gravity or radiation for explaining phenomena like human mobility. For validating such models, rich data sources such as mobile call records and social media data have recently become available. In the present work, I used mobile call records, billions of the messages of the online social network Twitter, and a detailed historical database of the voting data of United States counties for explaining aggregate patterns.

In Chapter 1, I introduced the literature on how human digital footprints can be used to model various real-world outcomes, especially by using data with geographical information. The background review was then narrowed down to the introduction of the Twitter social network and the potential limitations and biases present in the Twitter data, that is used in Chapters 3-5. Then, I decribed the formalization of the theory of urban scaling, and reviewed the most important empirical papers based on this theory, since the urban scaling methodology sets the background for Chapters 2 and 3 of the thesis.

In Chapter 2, I presented the results of a study on urban scaling in historical and recent presidential elections of the United States. Here, I showed that election results fit into the urban scaling framework and that a probabilistic model from economic complexity theory underlies the scaling phenomenon. In Chapter 3, I explored how the same urban scaling phenomenon is present in the word frequencies of the language people use on social media, namely the Twitter online social network. In this chapter, I also showed how qualitative linguistic laws, namely the Zipf's law and the Heaps law, hold in these online posts.

Chapter 4 analyzed the correlation of employment and unemployment rates of geographical areas in the United States with Twitter daily activity profiles. In this chapter, I also developed an algebraic approach for treating geographical areas in which the observed human activity timelines consist of the timelines of differently behaving groups of people. Chapter 5 investigated how spatiotemporal patterns in social media word frequencies predict mobile-phone based land use clustering in different cities. Finally, I summarized my findings in Chapter 6.

Összefoglalás

Disszertációmban emberi viselkedési mintázatokat próbáltam megérteni a komplex rendszerek vizsgálati módszereinek segítségével. Az irodalomban több példát is találhatunk arra, hogy fizikai modellek, mint például a gravitációs vonzás vagy a sugárzás, alkalmasak emberek kollektív viselkedésének leírására. Az ilyen és ehhez hasonló modellek alátámasztására ma már több különböző nagy adatforrás is a kutatók rendelkezésére áll. Doktori dolgozatomban ezek közül az adatforrások közül a Twitter szociális hálózat ingyenesen elérhető milliárdos nagyságrendű bejegyzéseit, mobiltelefonok hívásaiból származó adatokat, illetve egy részletes, hosszú időszakot felölelő amerikai elnökválasztási választási adatbázist használtam.

Az 1. fejezet található összefoglalóban először áttekintettem, milyen már létező lehetőségek vannak a digitális emberi lábnyomok felhasználására különböző szocioökonómiai mutatók modellezésében. A háttérirodalom taglalását ezek után leszűkítettem a Twitter online közösségi hálózatra, majd diszkutáltam, milyen lehetséges korlátjai vagy torzításai lehetnek a Twitterből származó adatoknak, mivel ezeket az adatokat használtam a disszertáció 3., 4. és 5. fejezetében. Majd az 1. fejezet második felében bevezettem a városi skálázás elméleti alapjait, illetve bemutattam az ehhez kapcsolódó legfontosabb empirikus tanulmányokat, melyek a 2. és a 3. fejezetekhez kapcsolódnak.

A 2. fejezetben megmutattam, hogy az Egyesült Államok városi területeinek szavazási eredményei több választásra visszamenőleg skálázási viselkedést mutatnak, és a skálázási paraméterek összefüggéseit egy komplexitási modell segítségével értelmeztem. A 3. fejezetben megvizsgáltam, hogy ugyanez a városi skálázási jelenség áll azon bejegyzések szógyakoriságainak hátterében, melyeket az emberek a Twitter online szociális hálózaton küldenek egymásnak. Ugyanebben a fejezetben megvizsgáltam az online bejegyzésekben két kvalitatív nyelvészeti törvény, a Zipf-törvény és a Heaps-törvény paramétereit.

A 4. fejezetben megmutattam, hogy egy adott földrajzi terület foglalkoztatottsági, illetve munkanélküliségi mutatói szignifikánsan korrelálnak az ugyanazon területen élő lakosság szociális hálózati adataiból általam újonnan bevezetett mérőszámmal. Egy lineáris algebrai modellel a megfigyelt átlagos napi aktivitást két különbözően viselkedő csoport napi aktivitására bontottam fel. Az 5. fejezetben pedig felügyelt gépi tanulás segítségével szociális hálózatból származó adatokból több különböző városban is rekonstruáltam a mobilhívási adatokon alapuló területhasználati klasztereket. Végül a 6. fejezetben foglaltam össze eredményeimet.

Publications

Publications supporting the thesis

In the order of Chapters 2-5.

1. Bokányi, E., Szállási, Z. & Vattay, G. "Universal scaling laws in metro area election results". *PLOS ONE* **13** (ed Braha, D.) e0192913 (2018)

2. Bokányi E., Kondor, D. & Vattay, G. "Scaling in words on Twitter". *submitted to Royal Society Open Science* (2019)

3. Bokányi, E., Lábszki, Z. & Vattay, G. "Prediction of employment and unemployment rates from Twitter daily rhythms in the US". *European Physical Journal Data Science* **6,** 14 (2017)

4. Bokányi E., Kallus, Zs. & Gódor, I. "Collective Sensing of Evolving Urban Structures: from Activity-based to Content-Aware Social Monitoring." *accepted in Environment and Planning B* (2019)

Other publications

5. Bokányi, E., Kondor, D., Dobos, L., Sebők, T., Stéger, J., Csabai, I. & Vattay, G. "Race, religion and the city: twitter word frequency patterns reveal dominant demographic dimensions in the United States". *Palgrave Communications* **2,** 16010 (2016)

6. Kmetty, Z., Koltai, J., Bokányi, E. & Bozsonyi, K. "Seasonality Pattern of Suicides in the US – a Comparative Analysis of a Twitter Based Bad-mood Index and Committed Suicides". *Intersections* **3** (2017)

7. Kallus, Z., Kondor, D., Stéger, J., Csabai, I., Bokányi, E. & Vattay, G. in *ICT Innovations 2017: Data-Driven Innovation. 9th International Conference, ICT Innovations 2017, Skopje, Macedonia, September 18-23, 2017, Proceedings* (eds Trajanov, D. & Bakeva, V.) 3–12 (Springer International Publishing, Cham, 2017)

8. Radnóczi, G., Bokányi, E., Erdélyi, Z. & Misják, F. "Size dependent spinodal decomposition in Cu-Ag nanoparticles". *Acta Materialia* **123,** 82–89 (2017)

9. Sóti, A., Bokányi, E. & Vattay, G. "Urban scaling of football followership on Twitter". *Acta Polytechnica Hungarica* **15,** 239–250 (2018)

Bibliography

10. Schweitzer, F. "Sociophysics". *Physics Today* **71,** 41–46 (2018).

11. Lazer, D. *et al.* "Computational Social Science". *Science* **323,** 721–723 (2009).

12. Mann, A. "Core Concept: Computational social science". *Proceedings of the National Academy of Sciences* **113,** 468–470 (2016).

13. Asur, S. & Huberman, B. A. "Predicting the Future with Social Media". in *International Conference on Web Intelligence and Intelligent Agent Technology* **1** (IEEE, 2010), 492–499. arXiv: 1003.5699.

14. Backstrom, L., Kleinberg, J., Kumar, R. & Novak, J. "Spatial variation in search engine queries". in *Proceeding of the 17th international conference on World Wide Web - WWW '08* (ACM Press, New York, New York, USA, 2008), 357.

15. Backstrom, L., Sun, E. & Marlow, C. "Find me if you can: improving geographical prediction with social and spatial proximity". in *Proceedings of the 19th international conference on World wide web* (2010), 61–70. arXiv: arXiv:1404.7152v1.

16. Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S. & Brilliant, L. "Detecting influenza epidemics using search engine query data." *Nature* **457,** 1012–1014 (2009).

17. Lazer, D., Kennedy, R., King, G. & Vespignani, A. "The Parable of Google Flu: Traps in Big Data Analysis". *Science* **343,** 1203–1205 (2014).

18. Ed de Quincey, Pantin, T., Theocharis, K. & Williams, N. "Potential of Social Media to Determine Hay Fever Seasons and Drug Efficacy". *Planet@ Risk* **2,** 293–297 (2014).

19. Paul, M. J. & Dredze, M. "You Are What You Tweet: Analyzing Twitter for Public Health". in *5th International AAAI Conference on Weblogs and Social Media* (2011), 265–272.

20. Prier, K. W., Smith, M. S., Giraud-Carrier, C. & Hanson, C. L. in *Social Computing, Behavioral-Cultural Modeling and Prediction* 18–25 (2011).

21. Lamb, A., Paul, M. J. & Dredze, M. "Separating Fact from Fear: Tracking Flu Infections on Twitter". in *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Association for Computational Linguistics, 2013), 789–795.

22. Dredze, M., Cheng, R., Paul, M. J. & Broniatowski, D. "HealthTweets.org : A Platform for Public Health Surveillance using Twitter". in *AAAI-14 Workshop on the World Wide Web and Public Health Intelligence* (2014), 2–3.

23. Paul, M. J. & Dredze, M. "Discovering health topics in social media using topic models". *PLoS ONE* **9** (2014).

24. Paul, M. J., Dredze, M., Broniatowski, D. A. & Generous, N. "Worldwide Influenza Surveillance through Twitter". in *AAAI Workshop - Technical Report* **WS-15-15** (2015), 6–11.

25. Pennebaker, J. W., Chung, C. K., Ireland, M., Gonzales, A. & Booth, R. J. "The Development and Psychometric Properties of LIWC2007". This article is published by LIWC Inc, Austin, Texas 78703 USA in conjunction with the LIWC2007 software program.

26. Chung, C. & Pennebaker, J. W. "The Psychological Functions of Function Words" (ed Fiedler, K.) 343–359 (Psychology Press, 2011).

27. Tausczik, Y. R. & Pennebaker, J. W. "The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods". *Journal of Language and Social Psychology* **29,** 24–54 (2010).

28. Pennebaker, J. J. & Chung, C. C. *Language and Social Dynamics* tech. rep. September (United States Army Research Institute for the Behavioral and Social Sciences, 2012).

29. Coppersmith, G. A., Harman, C. T. & Dredze, M. H. "Measuring Post Traumatic Stress Disorder in Twitter". in *In Proceedings of the 7th International AAAI Conference on Weblogs and Social Media (ICWSM).* **2** (2014), 23–45.

30. Reece, A. G., Reagan, A. J., Lix, K. L., Dodds, P. S., Danforth, C. M. & Langer, E. J. "Forecasting the onset and course of mental illness with Twitter data". *Scientific Reports* **7,** 1–23 (2017).

31. Eichstaedt, J. C. *et al.* "Psychological Language on Twitter Predicts County-Level Heart Disease Mortality". *Psychological Science* **26,** 159–169 (2015).

32. Surian, D., Nguyen, D. Q., Kennedy, G., Johnson, M., Coiera, E. & Dunn, A. G. "Characterizing Twitter Discussions About HPV Vaccines Using Topic Modeling and Community Detection". *Journal of Medical Internet Research* **18,** e232 (2016).

33. Beguerisse-Díaz, M., McLennan, A. K., Garduño-Hernández, G., Barahona, M. & Ulijaszek, S. J. "The 'who' and 'what' of #diabetes on Twitter". *Digital Health* **3,** 205520761668884 (2017).

34. Kalyanam, J., Velupillai, S., Doan, S., Conway, M. & Lanckriet, G. "Facts and Fabrications about Ebola: A Twitter Based Study". arXiv: `1508.02079` (2015).

35. Abbar, S., Mejova, Y. & Weber, I. "You Tweet What You Eat". in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems - CHI '15* (ACM Press, New York, New York, USA, 2015), 3197–3206. arXiv: `1412.4361`.

36. Widener, M. J. & Li, W. "Using geolocated Twitter data to monitor the prevalence of healthy and unhealthy food references across the US". *Applied Geography* **54,** 189–197 (2014).

37. Nguyen, Q. C. *et al.* "Leveraging geotagged Twitter data to examine neighborhood happiness, diet, and physical activity". *Applied Geography* **73,** 77–88 (2016).

38. Nguyen, Q. C. *et al.* "Twitter-derived neighborhood characteristics associated with obesity and diabetes". *Scientific Reports* **7,** 16425 (2017).

39. Eagle, N. & Pentland, A. "Reality mining: Sensing complex social systems". *Personal and Ubiquitous Computing* **10,** 255–268 (2006).

40. Crooks, A., Croitoru, A., Stefanidis, A. & Radzikowski, J. "#Earthquake: Twitter as a Distributed Sensor System". *Transactions in GIS* **17,** 124–147 (2013).

41. Sakaki, T., Okazaki, M. & Matsuo, Y. "Earthquake shakes Twitter users: real-time event detection by social sensors". in *Proceedings of the 19th International World Wide Web Conference* (2010), 851–860.

42. Blanford, J. I., Bernhardt, J. & Savelyev, A. "Tweeting and Tornadoes". in *The 11th International Conference on Information Systems for Crisis Response and Management* (2014), 319–323.

43. Shelton, T., Poorthuis, A., Graham, M. & Zook, M. "Mapping the data shadows of Hurricane Sandy: Uncovering the sociospatial dimensions of 'big data'". *Geoforum* **52,** 167–179 (2014).

44. Kryvasheyeu, Y., Chen, H., Obradovich, N., Moro, E., Van Hentenryck, P., Fowler, J. & Cebrian, M. "Rapid assessment of disaster damage using social media activity". *Science Advances* **2,** e1500779 (2016).

45. Graham, M. & Zook, M. "Augmented realities and uneven geographies: Exploring the geolinguistic contours of the web". *Environment and Planning A* **45,** 77–99 (2013).

46. Mislove, A., Lehmann, S., Ahn, Y.-y., Onnela, J.-P. & Rosenquist, J. N. "Understanding the Demographics of Twitter Users". *Artificial Intelligence,* 554–557 (2011).

47. Longley, P. A., Adnan, M. & Lansley, G. "The geotemporal demographics of Twitter usage". *Environment and Planning A* **47,** 465–484 (2015).

48. Longley, P. A. & Adnan, M. "Geo-temporal Twitter demographics". *International Journal of Geographical Information Science* **30,** 369–389 (2016).

49. Wood-Doughty, Z., Andrews, N., Marvin, R. & Dredze, M. "Predicting Twitter User Demographics from Names Alone". in *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media* (Association for Computational Linguistics, Stroudsburg, PA, USA, 2018), 105–111.

50. Mohammady, E. & Culotta, A. "Using county demographics to infer attributes of Twitter users". in *Joint Workshop on Social Dynamics and Personal Attributes in Social Media* (2014), 7–16.

51. Culotta, A., Ravi, N. K. & Cutler, J. "Predicting the Demographics of Twitter Users from Website Traffic Data number of unique neighbors count number of neighbor links". in *Proceedings of the International Conference on Web and Social Media (ICWSM), in press. Menlo Park, California: AAAI Press* (2015).

52. Nguyen, D., Gravel, R., Trieschnigg, D. & Meder, T. ""How old do you think I am?": A study of language and age in Twitter". in *Proceedings of the international AAAI conference on weblogs and social media* (2013), 439–448. arXiv: 1690219.1690245âĂŰ.

53. Pennacchiotti, M. & Popescu, A.-M. "Democrats, republicans and starbucks afficionados". in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '11* (ACM Press, New York, New York, USA, 2011), 430.

54. Pennacchiotti, M. & Popescu, A.-M. "A Machine Learning Approach to Twitter User Classification". in *Fifth International AAAI Conference on Weblogs and Social Media* (2011), 281–288.

55. Eagle, N., Macy, M. & Claxton, R. "Network Density and Economic Development". *Science* **328,** 1029–1031 (2010).

56. Eagle, N., de Montjoye, Y.-A. & Bettencourt, L. M. "Community Computing: Comparisons between Rural and Urban Societies Using Mobile Phone Data". in *2009 International Conference on Computational Science and Engineering* (IEEE, 2009), 144–150.

57. Wachs, J., Yasseri, T., Lengyel, B. & Kertész, J. "Social capital predicts corruption risk in towns". arXiv: 1810.05485 (2018).

58. Hawelka, B., Sitko, I., Beinat, E., Sobolevsky, S., Kazakopoulos, P. & Ratti, C. "Geo-located Twitter as proxy for global mobility patterns". *Cartography and Geographic Information Science* **41,** 260–271 (2014).

59. Messias, J., Benevenuto, F., Weber, I. & Zagheni, E. "From migration corridors to clusters: The value of Google+ data for migration studies". in *Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2016* (2016), 421–428. arXiv: 1607.00421.

60. Bokányi, E., Kondor, D., Dobos, L., Sebők, T., Stéger, J., Csabai, I. & Vattay, G. "Race, religion and the city: twitter word frequency patterns reveal dominant demographic dimensions in the United States". *Palgrave Communications* **2,** 16010 (2016).

61. Abitbol, J. L., Karsai, M., Magué, J.-P., Chevrot, J.-P. & Fleury, E. "Socioeconomic Dependencies of Linguistic Patterns in Twitter: A Multivariate Analysis". arXiv: 1804.01155 (2018).

62. Arnaboldi, M., Brambilla, M., Cassottana, B., Ciuccarelli, P. & Vantini, S. "How Twitter reveals Cities within Cities". in *Proceedings of the 7th 2016 International Conference on Social Media & Society - SMSociety '16* (ACM Press, New York, New York, USA, 2016), 1–11.

63. Lamanna, F., Lenormand, M., Salas-Olmedo, M. H., Romanillos, G., Gonçalves, B. & Ramasco, J. J. "Immigrant community integration in world cities". *PLOS ONE* **13** (ed Lambiotte, R.) e0191612 (2018).

64. *Twitter Company Metrics* tech. rep. (2018).

65. Osborne, M. & Dredze, M. "Facebook, Twitter and Google Plus for breaking news: Is there a winner?" in *Proceedings of the 8th International Conference on Weblogs and Social Media, ICWSM 2014* (2014), 611–614.

66. Kwak, H., Lee, C., Park, H. & Moon, S. "What is Twitter, a social network or a news media?" in *Proceedings of the 19th international conference on World wide web - WWW '10* (ACM Press, New York, New York, USA, 2010), 591. arXiv: 0809.1869v1.

67. Ch'ng, E. "Local Interactions and the Emergence of a Twitter Small-World Network". *Social Networking* **04,** 33–40 (2015).

68. Java, A., Song, X., Finin, T. & Tseng, B. "Why we twitter". in *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis - WebKDD/SNA-KDD '07* (ACM Press, New York, New York, USA, 2007), 56–65. arXiv: 1008.1253.

69. Bild, D. R., Liu, Y., Dick, R. P., Mao, Z. M. & Wallach, D. S. "Aggregate Characterization of User Behavior in Twitter and Analysis of the Retweet Graph". *ACM Transactions on Internet Technology* **15,** 1–24 (2015).

70. Stephens, M. & Poorthuis, A. "Follow thy neighbor: Connecting the social and the spatial networks on Twitter". *Computers, Environment and Urban Systems* **53,** 87–95 (2015).

71. Szüle, J., Kondor, D., Dobos, L., Csabai, I. & Vattay, G. "Lost in the City: Revisiting Milgram's Experiment in the Age of Social Networks". *PLoS ONE* **9** (ed Garcia-Ojalvo, J.) e111973 (2014).

72. Kallus, Z. *et al.* "Regional properties of global communication as reflected in aggregated Twitter data". in *2013 IEEE 4th International Conference on Cognitive Infocommunications (CogInfoCom)* (IEEE, 2013), 429–434. arXiv: 1311.1484.

73. Kallus, Z., Barankai, N., Szüle, J. & Vattay, G. "Spatial Fingerprints of Community Structure in Human Interaction Network for an Extensive Set of Large-Scale Regions". *PLOS ONE* **10** (ed Jiang, B.) e0126713 (2015).

74. Kallus, Z., Kondor, D., Stéger, J., Csabai, I., Bokányi, E. & Vattay, G. in *ICT Innovations 2017: Data-Driven Innovation. 9th International Conference, ICT Innovations 2017, Skopje, Macedonia, September 18-23, 2017, Proceedings* (eds Trajanov, D. & Bakeva, V.) 3–12 (Springer International Publishing, Cham, 2017).

75. Brockmann, D. & Helbing, D. "The hidden geometry of complex, network-driven contagion phenomena." *Science (New York, N.Y.)* **342,** 1337–1342 (2013).

76. Dobos, L. *et al.* "A multi-terabyte relational database for geo-tagged social network data". in *4th IEEE International Conference on Cognitive Infocommunications, CogInfoCom 2013 - Proceedings* (2013), 289–294.

77. Morstatter, F., Pfeffer, J., Liu, H. & Carley, K. M. "Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose". in *Association fo the Advanced of Artificial Intelligence* (2013). arXiv: 1306.5204.

78. Joseph, K., Landwehr, P. M. & Carley, K. M. in *Lecture Notes in Computer Science* 75–83 (2014).

79. Morstatter, F., Pfeffer, J. & Liu, H. "When is it biased? Assessing the Representativeness of Twitter's Streaming API". in *Proceedings of the 23rd International Conference on World Wide Web - WWW '14 Companion* (ACM Press, New York, New York, USA, 2014), 555–556. arXiv: 1401.7909.

80. Pfeffer, J., Mayer, K. & Morstatter, F. "Tampering with Twitter's Sample API". *EPJ Data Science* **7,** 50 (2018).

81. Duggan, M., Ellison, N. B., Lampe, C., Lenhart, A., Madden, M., Rainie, L. & Smith, A. *Pew Social Media Report 2015* tech. rep. January (Pew Research Center, 2015).

82. Hecht, B. & Stephens, M. "A Tale of Cities: Urban Biases in Volunteered Geographic Information". in *Proceedings of the International Workshop on Web and Social Media (ICWSM)* (2014).

83. Hargittai, E. & Litt, E. "The tweet smell of celebrity success: Explaining variation in Twitter adoption among a diverse group of young adults". *New Media & Society* **13,** 824–842 (2011).

84. Hofer, B., Lampoltshammer, T. J. & Belgiu, M. "Demography of Twitter users in the city of London". *Modern Trends in Cartography. Lecture Notes in Geoinformation and Cartography* (eds Brus, J., Vondrakova, A. & Vozenilek, V.) 199–211 (2015).

85. Sloan, L., Morgan, J., Burnap, P. & Williams, M. "Who Tweets? Deriving the Demographic Characteristics of Age, Occupation and Social Class from Twitter User Meta-Data". *Plos One* **10,** e0115545 (2015).

86. Oentaryo, R. J., Murdopo, A., Prasetyo, P. K. & Lim, E.-P. "On profiling bots in social media". in *International Conference on Social Informatics* (2016), 92–109. arXiv: 1609.00543.

87. Clark, E. M., Williams, J. R., Jones, C. A., Galbraith, R. A., Danforth, C. M. & Dodds, P. S. "Sifting robotic from organic text: A natural language approach for detecting automation on Twitter". *Journal of Computational Science* **16,** 1–7 (2016).

88. Kollanyi, B., Howard, P. N. & Woolley, S. C. *Bots and Automation over Twitter during the First U.S Presidential Debate* tech. rep. (2016), 1–5.

89. Howard, P., Bolsover, G. & Kollanyi, B. *Junk news and bots during the US election: What were Michigan voters sharing over Twitter* tech. rep. (2017), 1–5.

90. *United Nations, Department of Economic and Social Affairs. World's population increasingly urban with more than half living in urban areas* http://www.un.org/en/development/desa/news/population/world-urbanization-prospects-2014.html (2014).

91. Bettencourt, L. & West, G. "A unified theory of urban living". *Nature* **467,** 912–913 (2010).

92. Kleiber, M. "Body size and metabolic rate". *Physiological reviews* **27,** 511–541 (1947).

93. Bettencourt, L. M. A., Lobo, J., Helbing, D., Kuhnert, C. & West, G. B. "Growth, innovation, scaling, and the pace of life in cities". *Proceedings of the National Academy of Sciences* **104,** 7301–7306 (2007).

94. Bettencourt, L. & West, G. B. "Bigger Cities Do More with Less". *Scientific American* **305,** 52–53 (2011).

95. Arbesman, S., Kleinberg, J. M. & Strogatz, S. H. "Superlinear scaling for innovation in cities". *Physical Review E* **79,** 016115 (2009).

96. Bettencourt, L. M. A. "The origins of scaling in cities." *Science* **340,** 1438–1441 (2013).

97. Gomez-Lievano, A., Patterson-Lomba, O. & Hausmann, R. "Explaining the prevalence, scaling and variance of urban phenomena". *Nature Human Behaviour* **1,** 0012 (2016).

98. Hidalgo, C. A. & Hausmann, R. "The building blocks of economic complexity". *Proceedings of the National Academy of Sciences* **106,** 10570–10575 (2009).

99. Bettencourt, L. M. A., Lobo, J. & Youn, H. "The hypothesis of urban scaling: formalization, implications and challenges". arXiv: `1301.5919` (2013).

100. U.S. Census Bureau, D. I. S. "Metropolitan and Micropolitan Statistical Areas". *U.S. Census Bureau,* 2016 (2011).

101. Gomez-Lievano, A., Youn, H. & Bettencourt, L. M. A. "The statistics of urban scaling and their connection to Zipf's law". *PLoS ONE* **7** (2012).

102. Alves, L. G., Ribeiro, H. V., Lenzi, E. K. & Mendes, R. S. "Empirical analysis on the connection between power-law distributions and allometries for urban indicators". *Physica A: Statistical Mechanics and its Applications* **409,** 175–182 (2014).

103. Gabaix, X. "Power Laws in Economics and Finance". *Annual Review of Economics* **1,** 255–294 (2009).

104. Pumain, D., Paulus, F., Vacchiani-Marcuzzo, C. & Lobo, J. "An evolutionary theory for interpreting urban scaling laws". *Cybergeo* **2006,** 1–20 (2006).

105. Alves, L. G. A., Ribeiro, H. V., Lenzi, E. K. & Mendes, R. S. "Distance to the Scaling Law: A Useful Approach for Unveiling Relationships between Crime and Urban Metrics". *PLoS ONE* **8** (ed Perc, M.) e69580 (2013).

106. Alves, L. G. A., Mendes, R. S., Lenzi, E. K. & Ribeiro, H. V. "Scale-Adjusted Metrics for Predicting the Evolution of Urban Indicators and Quantifying the Performance of Cities". *PLOS ONE* **10** (ed Rozenblat, C.) e0134862 (2015).

107. Bettencourt, L. M. A., Lobo, J., Strumsky, D. & West, G. B. "Urban scaling and its deviations: Revealing the structure of wealth, innovation and crime across cities". *PLoS ONE* **5,** 20–22 (2010).

108. Sahasranaman, A. & Bettencourt, L. M. A. "Urban Geography and Scaling of Contemporary Indian Cities". arXiv: `1810.12004` (2018).

109. Shalizi, C. R. "Scaling and Hierarchy in Urban Economies". arXiv: `1102.4101` (2011).

110. Leitão, J. C., Miotto, J. M., Gerlach, M. & Altmann, E. G. "Is this scaling nonlinear?" *Royal Society Open Science* **3,** 1–11 (2016).

111. Stumpf, M. P. H. & Porter, M. A. "Critical Truths About Power Laws". *Science* **335** (2012).

112. Clauset, A., Shalizi, C. R. & Newman, M. E. J. "Power-Law Distributions in Empirical Data". *SIAM Review* **51,** 661–703 (2009).

113. Virkar, Y. & Clauset, A. "Power-law distributions in binned empirical data". *Annals of Applied Statistics* **8,** 89–119 (2014).

114. Filho, R. N. C., Almeida, M. P., Andrade, J. S. & Moreira, J. E. "Scaling behavior in a proportional voting process". *Physical Review E* **60,** 1067–1068 (1999).

115. Costa Filho, R. N., Almeida, M. P., Moreira, J. E. & Andrade, J. S. "Brazilian elections: Voting for a scaling democracy". *Physica A: Statistical Mechanics and its Applications* **322,** 698–700 (2003).

116. Lyra, M. L., Costa, U. M. S., Filho, R. N. C., Andrade, J. S. & Jr. "Generalized Zipf's Law in proportional voting processes". *Europhysics Letters* **62,** 5 (2002).

117. Fortunato, S. & Castellano, C. "Scaling and Universality in Proportional Elections". *Physical Review Letters* **99,** 138701 (2007).

118. Mantovani, M. C., Ribeiro, H. V., Moro, M. V., Picoli, S. & Mendes, R. S. "Scaling laws and universality in the choice of election candidates". *EPL (Europhysics Letters)* **96,** 48001 (2011).

119. Mantovani, M. C., Ribeiro, H. V., Lenzi, E. K., Picoli, S. & Mendes, R. S. "Engagement in the electoral processes: Scaling laws and the role of political positions". *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics* **88.** arXiv: 1308.2857 (2013).

120. Chatterjee, A., Mitrović, M. & Fortunato, S. "Universality in voting behavior: an empirical analysis". *Scientific Reports* **3,** 1049 (2013).

121. Klimek, P., Yegorov, Y., Hanel, R. & Thurner, S. "Statistical detection of systematic election irregularities". *Proceedings of the National Academy of Sciences* **109,** 16469–16473 (2012).

122. Bernardes, A., Stauffer, D. & Kertész, J. "Election results and the Sznajd model on Barabasi network". *The European Physical Journal B - Condensed Matter* **25,** 123–127 (2002).

123. Araújo, N. A., Andrade, J. S. & Herrmann, H. J. "Tactical voting in plurality elections". *PLoS ONE* **5,** 1–5 (2010).

124. Borghesi, C. & Bouchaud, J.-P. "Spatial correlations in vote statistics: a diffusive field model for decision-making". *The European Physical Journal B* **75,** 395–404 (2010).

125. Palombi, F. & Toti, S. "Stochastic Dynamics of the Multi-State Voter Model Over a Network Based on Interacting Cliques and Zealot Candidates". *Journal of Statistical Physics* **156,** 336–367 (2014).

126. Fernández-Gracia, J., Suchecki, K., Ramasco, J. J., San Miguel, M. & Eguíluz, V. M. "Is the Voter Model a Model for Voters?" *Physical Review Letters* **112,** 158701 (2014).

127. Borghesi, C., Raynal, J. C. & Bouchaud, J. P. "Election turnout statistics in many countries: Similarities, differences, and a diffusive field model for decision-making". *PLoS ONE* **7,** 1–12 (2012).

128. Braha, D. & de Aguiar, M. A. M. "Voting contagion: Modeling and analysis of a century of U.S. presidential elections". *PLOS ONE* **12** (ed Braunstein, L. A.) e0177970 (2017).

129. Florida, R. *Mapping How America's Metro Areas Voted: The geography of the 2016 election is spiky* http://www.citylab.com/politics/2016/12/mapping-how-americas-metro-areas-voted/508313/ (2016).

130. Florida, R. "The Rise of the Creative Class–Revisited: Revised and Expanded" (Basic books, 2014).

131. Goodwin, M. J. & Heath, O. "The 2016 Referendum, Brexit and the Left Behind: An Aggregate-level Analysis of the Result". *The Political Quarterly* **87,** 323–332 (2016).

132. Leip, D. "Dave Leip U.S. Presidential General County Election Results" (Harvard Dataverse, 2016).

133. *Electoral Commission* http://electoralcommission.org.uk (2016).

134. *Metropolitan and Micropolitan Statistical Areas Main - US Census Bureau 2016* https://www.census.gov/population/metro/ (2016).

135. *CMS's SSA to FIPS CBSA and MSA County Crosswalk* http://www.nber.org/data/cbsa-msa-fips-ssa-county-crosswalk.html (2017).

136. *United Kingdom: Countries and Major Cities - Population Statistics in Maps and Charts* https://www.citypopulation.de/UK-Cities.html (2016).

137. Lobo, J., Bettencourt, L. M. A., Strumsky, D. & West, G. B. "Urban Scaling and the Production Function for Cities". *PLoS ONE* **8** (ed Hidalgo, C. A.) e58407 (2013).

138. McCarty, N., Poole, K. T. & Rosenthal, H. "Polarized America: The Dance of Ideology and Unequal Riches", 642. arXiv: arXiv:1011.1669v3 (MIT Press, 2008).

139. Rabinowitz, G. & Macdonald, S. E. "A Directional Theory of Issue Voting". *The American Political Science Review* **83,** 93–121 (1989).

140. Denver, D. & Hands, G. "Issues, principles or ideology? How young voters decide". *Electoral Studies* **9,** 19–36 (1990).

141. Crystal, D. "Internet linguistics: A student guide" (Routledge, 2011).

142. Altmann, E. G. & Gerlach, M. in *Creativity and Universality in Language* 7–26 (2016). arXiv: 1502.03296.

143. Gerlach, M. & Altmann, E. G. "Scaling laws and fluctuations in the statistics of word frequencies". *New Journal of Physics* **16,** 113010 (2014).

144. Altmann, E. G., Pierrehumbert, J. B. & Motter, A. E. "Beyond word frequency: Bursts, lulls, and scaling in the temporal distributions of words". *PLoS ONE* **4** (ed Scalas, E.) e7678 (2009).

145. Goel, R., Soni, S., Goyal, N., Paparrizos, J., Wallach, H., Diaz, F. & Eisenstein, J. in *Social Informatics* 41–57 (2016). arXiv: 1609.02075.

146. Gonçalves, B., Loureiro-Porto, L., Ramasco, J. J. & Sánchez, D. "Mapping the Americanization of English in space and time". *PLOS ONE* **13** (ed Preis, T.) e0197741 (2018).

147. Wang, W., Chen, L., Thirunarayan, K. & Sheth, A. P. "Cursing in English on twitter". in *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing - CSCW '14* (ACM Press, New York, New York, USA, 2014), 415–425.

148. Gauthier, M., Guille, A., Rico, F. & Deseille, A. "Text mining and Twitter to analyze British swearing habits". in *Proceedings of International Conference on Twitter for Research* (2015), 28–42. arXiv: arXiv:1011.1669v3.

149. Byrne, E. & Corney, D. "Sweet FA: Sentiment, Swearing and Soccer". in *SoMuS@ICMR* (2014).

150. Blodgett, S. L., Green, L. & O'Connor, B. "Demographic Dialectal Variation in Social Media: A Case Study of African-American English". arXiv: 1608.08868 (2016).

151. Cheng, Z., Caverlee, J. & Lee, K. "You are where you tweet: A content-based approach to geo-locating Twitter users". in *Proceedings of the 19th ACM International Conference on Information and Knowledge Management* (2010), 759–768.

152. Ferrara, E., Varol, O., Menczer, F. & Flammini, A. "Traveling trends". in *Proceedings of the first ACM conference on Online social networks - COSN '13* (ACM Press, New York, New York, USA, 2013), 213–222. arXiv: 1310.2671.

153. Eisenstein, J., O'Connor, B., Smith, N. A. & Xing, E. P. "Diffusion of Lexical Change in Social Media". *PLoS ONE* **9,** 1–13 (2012).

154. Mitchell, L., Frank, M. R., Harris, K. D., Dodds, P. S. & Danforth, C. M. "The geography of happiness: connecting twitter sentiment and expression, demographics, and objective characteristics of place." *PloS one* **8,** e64417 (2013).

155. Arcaute, E., Hatna, E., Ferguson, P., Youn, H., Johansson, A. & Batty, M. "Constructing cities , deconstructing scaling laws". *Journal of The Royal Society Interface* **12,** 3–6 (2015).

156. Cottineau, C., Hatna, E., Arcaute, E. & Batty, M. "Diverse cities or the systematic paradox of Urban Scaling Laws". *Computers, Environment and Urban Systems* **63,** 80–94 (2017).

157. Yakubo, K., Saijo, Y. & Korošak, D. "Superlinear and sublinear urban scaling in geographical networks modeling cities". *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics* **90,** 1–10 (2014).

158. Schläpfer, M. *et al.* "The scaling of human interactions with city size". *Journal of the Royal Society, Interface / the Royal Society* **11,** 20130789– (2014).

159. Youn, H., Bettencourt, L. M. A., Lobo, J., Strumsky, D., Samaniego, H. & West, G. B. "Scaling and universality in urban economic diversification". *Journal of The Royal Society Interface* **13,** 20150937 (2016).

160. Bokányi, E., Szállási, Z. & Vattay, G. "Universal scaling laws in metro area election results". *PLoS ONE* **13** (2018).

161. Schläpfer, M., Lee, J. & Bettencourt, L. M. A. "Urban Skylines: building heights and shapes as measures of city size", 1–17 (2015).

162. Oliveira, M., Bastos-Filho, C. & Menezes, R. "The scaling of crime concentration in cities". *PLoS ONE* **12** (2017).

163. Hanley, Q. S., Lewis, D. & Ribeiro, H. V. "Rural to urban population density scaling of crime and property transactions in english and welsh parliamentary constituencies". *PLoS ONE* **11,** 25–27 (2016).

164. Bojic, I., Belyi, A., Ratti, C. & Sobolevsky, S. "Scaling of foreign attractiveness for countries and states". *Applied Geography* **73,** 47–52 (2016).

165. Zipf, G. K. "Selected studies of the principles of relative frequency in language" (Harvard University Press, 1932).

166. Takahashi, S. & Tanaka-Ishii, K. "Assessing Language Models with Scaling Properties". arXiv: 1804.08881 (2018).

167. Szalay, A. S., Gray, J., Fekete, G., Kunszt, P. Z., Kukol, P. & Thakar, A. "Indexing the Sphere with the Hierarchical Triangular Mesh". arXiv: 0701164 [cs] (2007).

168. Kondor, D., Dobos, L., Csabai, I., Bodor, A., Vattay, G., Budavári, T. & Szalay, A. S. "Efficient classification of billions of points into complex geographic regions using hierarchical triangular mesh". in *Proceedings of the 26th International Conference on Scientific and Statistical Database Management - SSDBM '14* (ACM Press, New York, New York, USA, 2014), 1–4.

169. *Global Administrative Areas* http://gadm.org (2016).

170. US Census Bureau. *Metropolitan and Micropolitan Statistical Areas Totals: 2010-2017* `https://www.census.gov/data/tables/2017/demo/popest/total-metro-and-micro-statistical-areas.html` (2017).

171. Ferrer i Cancho, R. & Solé, R. V. "Two Regimes in the Frequency of Words and the Origins of Complex Lexicons: Zipf's Law Revisited". *Journal of Quantitative Linguistics* **8,** 165–173 (2001).

172. Alstott, J., Bullmore, E. & Plenz, D. "Powerlaw: A python package for analysis of heavy-tailed distributions". *PLoS ONE* **9** (2014).

173. Goldstein, M. L., Morris, S. A. & Yen, G. G. "Problems with fitting to the power distribution". *Eur. Phys. J. B* **41,** 255–258 (2004).

174. *Ranking the Latino population in metropolitan areas | Pew Research Center* `http://www.pewhispanic.org/2016/09/08/5-ranking-the-latino-population-in-metropolitan-areas/` (2018).

175. Ferrer i Cancho, R. "The variation of Zipf's law in human language". *European Physical Journal B* **44,** 249–257 (2005).

176. Ferrer i Cancho, R. & Solé, R. V. "Least effort and the origins of scaling in human language". *Proceedings of the National Academy of Sciences* **100,** 788–791 (2003).

177. Crystal, D. "Texting". *ELT Journal* **62,** 77–83 (2008).

178. Montemurro, M. A. "Beyond the Zipf-Mandelbrot law in quantitative linguistics". *Physica A: Statistical Mechanics and its Applications* **300,** 567–578 (2001).

179. Petersen, A. M., Tenenbaum, J. N., Havlin, S., Stanley, H. E. & Perc, M. "Languages cool as they expand: Allometric scaling and the decreasing need for new words". *Scientific Reports* **2,** 18–20 (2012).

180. Aschoff, J. & Wever, R. "Human circadian rhythms: a multioscillatory system". *Federation proceedings* **35,** 236–232 (1976).

181. Cagnacci, A., Elliott, J. A. & Yen, S. S. "Melatonin: a major regulator of the circadian rhythm of core temperature in humans." *The Journal of Clinical Endocrinology & Metabolism* **75,** 447–452 (1992).

182. Refinetti, R. & Menaker, M. "The circadian rhythm of body temperature". *Physiology & Behavior* **51,** 613–637 (1992).

183. Cajochen, C., Kräuchi, K. & Wirz-Justice, A. "Role of Melatonin in the Regulation of Human Circadian Rhythms and Sleep". *Journal of Neuroendocrinology* **15,** 432–437 (2003).

184. Taillard, J., Philip, P. & Bioulac, B. "Morningness/eveningness and the need for sleep". *Journal of Sleep Research* **8,** 291–295 (1999).

185. Aledavood, T., Lehmann, S. & Saramäki, J. "Digital daily cycles of individuals". *Frontiers in Physics* **3,** 1–7 (2015).

186. Saramäki, J. & Moro, E. "From seconds to months: an overview of multi-scale dynamics of mobile telephone calls". *The European Physical Journal B* **88,** 1–10 (2015).

187. Reades, J., Calabrese, F., Sevtsuk, a. & Ratti, C. "Cellular Census". *Pervasive computing* **6,** 30–38 (2007).

188. Reades, J., Calabrese, F. & Ratti, C. "Eigenplaces: Analysing cities using the space - Time structure of the mobile phone network". *Environment and Planning B: Planning and Design* **36,** 824–836 (2009).

189. Calabrese, F., Reades, J. & Ratti, C. "Eigenplaces : Segmenting Space through Digital Signatures Eigenplaces : Segmenting Space through Digital Signatures". *IEEE Pervasive Computing* **9,** 78–84 (2010).

190. Soto, V. & Frías-Martínez, E. "Automated land use identification using cell-phone records". in *Proceedings of the 3rd ACM international workshop on MobiArch - HotPlanet '11* (2011), 17.

191. Toole, J. L., Lin, Y.-R., Muehlegger, E., Shoag, D., González, M. C. & Lazer, D. "Tracking employment shocks using mobile phone data." *Journal of the Royal Society, Interface / the Royal Society* **12,** 20150185 (2015).

192. Pei, T., Sobolevsky, S., Ratti, C., Shaw, S.-L., Li, T. & Zhou, C. "A new insight into land use classification based on aggregated mobile phone data". *International Journal of Geographical Information Science* **28,** 1988–2007 (2014).

193. Louail, T. *et al.* "Uncovering the spatial structure of mobility networks." *Nature communications* **6,** 6007 (2015).

194. Grauwin, S., Sobolevsky, S., Moritz, S., Gódor, I. & Ratti, C. in *Computational Approaches for Urban Environments* 363–387 (Springer International Publishing, Cham, 2015).

195. Kondor, D., Thebault, P., Grauwin, S., Gódor, I., Moritz, S., Sobolevsky, S. & Ratti, C. "Visualizing signatures of human activity in cities across the globe". *arXiv preprint* **3,** 1–6 (2015).

196. Lenormand, M. *et al.* "Comparing and modelling land use organization in cities". *Royal Society Open Science* **2,** 150449 (2015).

197. Cici, B., Gjoka, M., Markopoulou, A. & Butts, C. T. "On the Decomposition of Cell Phone Activity Patterns and their Connection with Urban Ecology." in *Proceedings of ACM MobiHoc '15* (eds Shen, S., Sun, Y., Chen, J., Zhang, J. & Zussman, G.) (ACM, 2015), 317–326.

198. Candia, J., González, M. C., Wang, P., Schoenharl, T., Madey, G. & Barabási, A.-L. "Uncovering individual and collective human dynamics from mobile phone records". *J. Phys. A: Math. Theor* **41,** 224015–11 (2008).

199. Douglass, R. W., Meyer, D. A., Ram, M., Rideout, D. & Song, D. "High resolution population estimates from telecommunications data". *EPJ Data Science* **4,** 1–13 (2014).

200. Schneider, C. M., Belik, V., Couronne, T., Smoreda, Z. & Gonzalez, M. C. "Unravelling Daily Human Mobility Motifs". *Journal of The Royal Society Interface* **10,** 20130246(1–8) (2013).

201. Smith-Clarke, C., Mashhadi, A. & Capra, L. "Poverty on the cheap: Estimating Poverty Maps Using Aggregated Mobile Communication Networks". in *Proceedings of the 32nd annual ACM conference on Human factors in computing systems - CHI '14* (ACM Press, New York, New York, USA, 2014), 511–520.

202. Bogomolov, A., Lepri, B., Staiano, J., Oliver, N., Pianesi, F. & Pentland, A. "Once Upon a Crime". in *Proceedings of the 16th International Conference on Multimodal Interaction - ICMI '14* (ACM Press, New York, New York, USA, 2014), 427–434. arXiv: 1409.2983.

203. Jiang, S., Ferreira, J. & González, M. C. "Clustering daily patterns of human activities in the city". *Data Mining and Knowledge Discovery* **25,** 478–510 (2012).

204. Ettredge, B. M., Gerdes, J. & Karuga, G. "Using web-based search data to predict macroeconomic statistics". *Communications of the ACM* **48,** 87–92 (2005).

205. Choi, H. & Varian, H. "Predicting the Present with Google Trends". *Economic Record* **88,** 2–9 (2012).

206. Pavlicek, J. & Kristoufek, L. "Nowcasting Unemployment Rates with Google Searches: Evidence from the Visegrad Group Countries". *Plos One* **10,** e0127084 (2015).

207. Proserpio, D., Counts, S. & Jain, A. "The psychology of job loss: using social media data to characterize and predict unemployment." in *WebSci* (eds Nejdl, W., Hall, W., Parigi, P. & Staab, S.) (ACM, 2016), 223–232.

208. Toole, J. L., Lin, Y.-R., Muehlegger, E., Shoag, D., González, M. C. & Lazer, D. "Tracking employment shocks using mobile phone data." *Journal of the Royal Society, Interface / the Royal Society* **12,** 20150185– (2015).

209. Llorente, A., Garcia-Herranz, M., Cebrian, M. & Moro, E. "Social Media Fingerprints of Unemployment". *PLOS ONE* **10** (ed Moreno, Y.) e0128692 (2015).

210. *United States Census 2010* `http://www2.census.gov/census_2010/` (2016).

211. *Local Area Unemployment Statistics* `http://www.bls.gov/lau/` (2016).

212. Goodchild, M. F. "Citizens as sensors: the world of volunteered geography". *GeoJournal* **69,** 211–221 (2007).

213. Dashdorj, Z. & Sobolevsky, S. in *Transactions on Large-Scale Data- and Knowledge-Centered Systems XXVII* 159–176 (2015). arXiv: `1510.02995`.

214. Zanini, P., Shen, H. & Truong, Y. "Understanding resident mobility in Milan through independent component analysis of Telecom Italia mobile usage data". *The Annals of Applied Statistics* **10,** 812–833 (2016).

215. Ríos, S. A. & Muñoz, R. "Land Use detection with cell phone data using topic models: Case Santiago, Chile". *Computers, Environment and Urban Systems* **61,** 39–48 (2017).

216. Rubio, A., Sanchez, A. & Frias-Martinez, E. "Adaptive non-parametric identification of dense areas using cell phone records for urban analysis". *Engineering Applications of Artificial Intelligence* **26,** 551–563 (2013).

217. Li, M., Shen, Z. & Hao, X. "Revealing the relationship between spatio-temporal distribution of population and urban function with social media data". *GeoJournal* **81,** 919–935 (2016).

218. Cáceres, R., Rowland, J., Small, C. & Urbanek, S. "Exploring the Use of Urban Greenspace through Cellular Network Activity". in *Proc. of 2nd Workshop on Pervasive Urban Applications (PURBA)* (2012), 1–8.

219. Frias-Martinez, V., Soto, V., Hohwald, H. & Frias-Martinez, E. "Characterizing Urban Landscapes Using Geolocated Tweets". in *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing* (IEEE, 2012), 239–248.

220. Frias-Martinez, V. & Frias-Martinez, E. "Spectral clustering for sensing urban land use using Twitter activity". *Engineering Applications of Artificial Intelligence* **35,** 237–245 (2014).

221. Steiger, E., Resch, B. & Zipf, A. "Exploration of spatiotemporal and semantic clusters of Twitter data using unsupervised neural networks". *International Journal of Geographical Information Science* **30,** 1694–1716 (2016).

222. Picornell, M., Ruiz, T., Lenormand, M., Ramasco, J. J., Dubernet, T. & Frías-Martínez, E. "Exploring the potential of phone call data to characterize the relationship between social network and travel behavior". *Transportation* **42,** 647–668 (2015).

223. García-Palomares, J. C., Salas-Olmedo, M. H., Moya-Gómez, B., Condeço-Melhorado, A. & Gutiérrez, J. "City dynamics through Twitter: Relationships between land use and spatiotemporal demographics". *Cities* **72,** 310–319 (2018).

224. Kling, F. & Pozdnoukhov, A. "When a city tells a story". in *Proceedings of the 20th International Conference on Advances in Geographic Information Systems - SIGSPATIAL '12* (ACM Press, New York, New York, USA, 2012), 482.

225. Jiang, S., Fiore, G. A., Yang, Y., Ferreira, J., Frazzoli, E. & González, M. C. "A review of urban computing for mobile phone traces". in *Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing - UrbComp '13* (ACM Press, New York, New York, USA, 2013), 1.

226. Blondel, V. D., Decuyper, A. & Krings, G. "A survey of results on mobile phone datasets analysis". *EPJ Data Science* **4,** 10 (2015).

227. Picornell, M. *et al.* "Cross-Checking Different Sources of Mobility Information". *PLoS ONE* **9** (ed Moreno, Y.) e105184 (2014).

228. O'Connor, B., Krieger, M. & Ahn, D. "TweetMotif : Exploratory search and topic summarization for Twitter". in *4th International AAAI Conference on Weblogs and Social Media* (2010), 2–3.

229. *hunspell 0.5.0 : Python Package Index* `https://pypi.python.org/pypi/hunspell` (2017).

230. Nemeth, L., Tron, V., Halacsy, P., Kornai, A., Rung, A. & Shakadat, I. "Leveraging the open source ispell codebase for minority language analysis". in *Proceedings of SALTMIL* (2004), 56–59.

231. *Stopwords* `http://www.ranks.nl/stopwords` (2017).

232. Liaw, A. & Wiener, M. "Classification and Regression by randomForest". *R News* **2,** 18–22 (2002).

233. Hanley, A. & McNeil, J. "The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve". *Radiology* **143,** 29–36 (1982).

234. Fawcett, T. "An introduction to ROC analysis". *Pattern Recognition Letters* **27,** 861–874 (2006).

235. Cai, G., Lee, K. & Lee, I. in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 320–332 (2016).

236. Soltani, K., Soliman, A., Padmanabhan, A. & Wang, S. "UrbanFlow". in *Proceedings of the XSEDE16 on Diversity, Big Data, and Science at Scale - XSEDE16* (ACM Press, New York, New York, USA, 2016), 1–8.

237. Cesare, N., Grant, C. & Nsoesie, E. O. "Detection of User Demographics on Social Media: A Review of Methods and Recommendations for Best Practices". *CoRR,* 1–18 (2017).

List of Figures

List of Tables

ADATLAP

a doktori értekezés nyilvánosságra hozatalához*

I. A doktori értekezés adatai

A szerző neve: Bokányi Eszter

MTMT-azonosító: 10055125

A doktori értekezés címe és alcíme: Analysis of complex socio-economic systems

DOI-azonosító: 10.15476/ELTE.2019.054

A doktori iskola neve: Fizika Doktori Iskola

A doktori iskolán belüli doktori program neve: Statisztikus Fizika, Biológiai Fizika és Kvantumrendszerek Fizikája program

A témavezető neve és tudományos fokozata: Vattay Gábor, DSc, egyetemi tanár

A témavezető munkahelye: ELTE TTK Komplex Rendszerek Fizikája Tanszék

II. Nyilatkozatok

1. A doktori értekezés szerzőjeként

a) hozzájárulok, hogy a doktori fokozat megszerzését követően a doktori értekezésem és a tézisek nyilvánosságra kerüljenek az ELTE Digitális Intézményi Tudástárban. Felhatalmazom a Természettudományi kar Dékáni Hivatal Doktori, Habilitációs és Nemzetközi Ügyek Csoportjának ügyintézőjét, hogy az értekezést és a téziseket feltöltse az ELTE Digitális Intézményi Tudástárba, és ennek során kitöltse a feltöltéshez szükséges nyilatkozatokat.

~~b) kérem, hogy a mellékelt kérelemben részletezett szabadalmi, illetőleg oltalmi bejelentés közzétételéig a doktori értekezést ne bocsássák nyilvánosságra az Egyetemi Könyvtárban és az ELTE Digitális Intézményi Tudástárban;~~

~~c) kérem, hogy a nemzetbiztonsági okból minősített adatot tartalmazó doktori értekezést a minősítés (*dátum*)-ig tartó időtartama alatt ne bocsássák nyilvánosságra az Egyetemi Könyvtárban és az ELTE Digitális Intézményi Tudástárban;~~

~~d) kérem, hogy a mű kiadására vonatkozó mellékelt kiadó szerződésre tekintettel a doktori értekezést a könyv megjelenéséig ne bocsássák nyilvánosságra az Egyetemi Könyvtárban, és az ELTE Digitális Intézményi Tudástárban csak a könyv bibliográfiai adatait tegyék közzé. Ha a könyv a fokozatszerzést követően egy évig nem jelenik meg, hozzájárulok, hogy a doktori értekezésem és a tézisek nyilvánosságra kerüljenek az Egyetemi Könyvtárban és az ELTE Digitális Intézményi Tudástárban.~~

2. A doktori értekezés szerzőjeként kijelentem, hogy

a) az ELTE Digitális Intézményi Tudástárba feltöltendő doktori értekezés és a tézisek saját eredeti, önálló szellemi munkám és legjobb tudomásom szerint nem sértem vele senki szerzői jogait;

b) a doktori értekezés és a tézisek nyomtatott változatai és az elektronikus adathordozón benyújtott tartalmak (szöveg és ábrák) mindenben megegyeznek.

3. A doktori értekezés szerzőjeként hozzájárulok a doktori értekezés és a tézisek szövegének plágiumkereső adatbázisba helyezéséhez és plágiumellenőrző vizsgálatok lefuttatásához.

Kelt: 2019. március 7.

Bokányi Eszter

.....

a doktori értekezés szerzőjének aláírása

**ELTE SZMSZ SZMR 12. sz. melléklet*